Journal of applied logics (Print) ISSN 2631-9810
Journal of applied logics (Online) ISSN 2631-9829

Journal of Applied Logics

The IfCoLog Journal of Logics and their Applications

Volume 6 ● Issue 3 ● May 2019

Available online at
www.collegepublications.co.uk/journals/ifcolog/
Free open access

Volume 6 ● Issue 3 ● May 2019

Journal of Applied Logics The IfCoLog Journal of Logics and their Applications

Contents

Articles

Sponsored by
ELSEVIER

ISBN 978-1-84890-305-0
90000
9 781848 903050

Published by
IFCoLog

9 781848 903050

7.44 x 9.69
246 mm x 189 mm

.396
10.05mm

7.44 x 9.69
246 mm x 189 mm

Content Type: Black & White
Paper Type: White
Page Count: 186
File Type: InDesign
Request ID: CSS2612877

Perfect Bound Cover Template

Lightning Source

Document Size: 19" x 12"
305 x 483mm

Journal of Applied Logics - IfCoLog Journal of Logics and their Applications

Volume 6, Number 3

May 2019

Disclaimer
Statements of fact and opinion in the articles in Journal of Applied Logics - IfCoLog Journal of Logics and their Applications (JALs-FLAP) are those of the respective authors and contributors and not of the JALs-FLAP. Neither College Publications nor the JALs-FLAP make any representation, express or implied, in respect of the accuracy of the material in this journal and cannot accept any legal responsibility or liability for any errors or omissions that may be made. The reader should make his/her own evaluation as to the appropriateness or otherwise of any experimental technique described.

ISBN 978-1-84890-305-0
ISSN (E) 2631-9829
ISSN (P) 2631-9810

College Publications
Scientific Director: Dov Gabbay
Managing Director: Jane Spurr

http://www.collegepublications.co.uk

Editorial Board

Scope and Submissions

This journal considers submission in all areas of pure and applied logic, including:

pure logical systems
proof theory
constructive logic
categorical logic
modal and temporal logic
model theory
recursion theory
type theory
nominal theory
nonclassical logics
nonmonotonic logic
numerical and uncertainty reasoning
logic and AI
foundations of logic programming
belief change/revision
systems of knowledge and belief
logics and semantics of programming
specification and verification
agent theory
databases
dynamic logic
quantum logic
algebraic logic
logic and cognition
probabilistic logic
logic and networks
neuro-logical systems
complexity
argumentation theory
logic and computation
logic and language
logic engineering
knowledge-based systems
automated reasoning
knowledge representation
logic in hardware and VLSI
natural language
concurrent computation
planning

This journal will also consider papers on the application of logic in other subject areas: philosophy, cognitive science, physics etc. provided they have some formal content.

Submissions should be sent to Jane Spurr (jane.spurr@kcl.ac.uk) as a pdf file, preferably compiled in LaTeX using the IFCoLog class file.

Contents

ARTICLES

New Foundations for Imperative Logic IV: Natural Deduction

Peter B. M. Vranas
University of Wisconsin-Madison, USA
vranas@wisc.edu

Abstract

Sentential Pure Imperative Logic (SPIL) deals with arguments from imperative premises to imperative conclusions (i.e., pure imperative arguments) that do not contain quantifiers or modal operators. I introduce a formal language and a natural deduction system for SPIL. I provide the formal language with a semantics, and I prove that the natural deduction system is sound and complete with respect to that semantics.

1 Introduction

In this paper, I present a sound and complete natural deduction system for *Sentential Pure Imperative Logic* (SPIL), which deals with arguments from imperative premises to imperative conclusions but does not include quantifiers or modal operators. I provide an imperative formal language, as well as replacement and inference rules that can be used to derive a conclusion from a set of premises. The replacement and inference rules are intended to represent natural patterns of reasoning, but their justification is not limited to intuitions about naturalness. The justification relies crucially on the result—which I prove—that derivability by those rules corresponds to a semantic definition of argument validity that I have developed at length in previous papers ([10, 12]; see also [8, 9, 11]) and that I develop further

I am grateful to John Mackay, Michael Titelbaum, Berislav Žarnić, several anonymous reviewers, and especially Aviv Hoffmann and an editor of the *Journal of Applied Logics* for comments, and to Jeremy Avigad, David Makinson, and especially Jörg Hansen for help. Thanks also to Fabrizio Cariani, Hannah Clark-Younger, Kit Fine, Malcolm Forster, Casey Hart, Daniel Hausman, Blake Myers, David O'Brien, Brian Skyrms, and Elliott Sober for interesting questions, and to my mother and Jane Spurr for typing the bulk of the paper. Material from this paper was presented at the University of Wisconsin-Madison (Department of Mathematics, April 2014, and Department of Philosophy, May 2014), the Madison Informal Formal Epistemology Meeting (April 2014), the 12th International Conference on Deontic Logic and Normative Systems (DEON 2014), the New York University Workshop "Imperatives and Deontic Modals" (March 2016), and the 13th International Conference on Deontic Logic and Normative Systems (DEON 2016).

here by introducing *interpretations* of imperative formal languages. I do not presuppose any familiarity with the previous papers.[1]

2 Syntax

The (imperative formal) language of SPIL has the following symbols: the connectives '$\sim$', '&', '$\vee$', '$\rightarrow$', and '$\leftrightarrow$', the punctuation symbols '(' and ')', the *imperative operator* '!' ("let it be the case that"), and the (infinitely many) sentence letters 'A', 'B', ..., 'Z', 'A'', 'B'', ..., 'Z'', 'A''', 'B''', ... (One could also define languages of SPIL with different sentence letters or with only finitely many sentence letters, but for simplicity I define only a single language of SPIL.) The *declarative sentences* of SPIL can be built up from sentence letters as in classical sentential logic. The *imperative sentences* of SPIL are all and only those finite strings of symbols (understood as ordered n-tuples of symbols) that can be built up from declarative sentences by applying the following formation rules (R1 must be applied at least once):

(R1) If p is a declarative sentence, then $\ulcorner !p \urcorner$ is an imperative sentence.

(R2) If i and j are imperative sentences, then $\ulcorner \sim i \urcorner$, $\ulcorner (i \;\&\; j) \urcorner$, and $\ulcorner (i \vee j) \urcorner$ are also imperative sentences.

(R3) If p is a declarative sentence and i is an imperative sentence, then $\ulcorner (p \rightarrow i) \urcorner$, $\ulcorner (i \rightarrow p) \urcorner$, $\ulcorner (p \leftrightarrow i) \urcorner$, and $\ulcorner (i \leftrightarrow p) \urcorner$ are imperative sentences.

A *sentence* (of SPIL) is either a declarative sentence or an imperative sentence. It follows from these definitions that a sentence is imperative iff it contains at least one occurrence of '!' and is declarative iff it contains no occurrence of '!' (so no sentence is both declarative and imperative). Throughout this paper, I use the following notation: (1) φ and ψ are (declarative or imperative) sentences, (2) $p, q, r, p', \ldots$ are declarative sentences, (3) $i, j, k, i', \ldots$ are imperative sentences, and (4) e is a sentence letter. For simplicity, I usually omit outermost parentheses.

[1] There is hardly any previous work on this subject. To my knowledge, only two logic textbooks cover symbolization of imperative English sentences and natural deduction for imperative logic: [2] (a descendant of [1]) and [6] (a descendant of [4]; see also [5, pp. 181–6]. These textbooks, however, rely on inadequate definitions of validity (see [10, 12]). Relying on my definition of validity for arguments with only imperative premises and conclusions [10], Hansen [7] has provided sound and complete sets of inference rules for a formal language with only one imperative connective. See also [3, pp. 625–6].

3 Semantics

An *interpretation* of the language of SPIL is an ordered pair $m = \langle \mathbb{S}, \mathbb{F} \rangle$, where $\mathbb{S}$ is a set of sentence letters and $\mathbb{F}$ is a *favoring relation*, namely a three-place relation on declarative sentences that satisfies two conditions. First, the *intensionality condition*: for any p, q, and r and any p', q', and r' interderivable in classical sentential logic with p, q, and r respectively, $\langle p, q, r \rangle \in \mathbb{F}$ iff $\langle p', q', r' \rangle \in \mathbb{F}$. Second, the *asymmetry condition*: for any p, q, and r, it is not the case that both $\langle p, q, r \rangle \in \mathbb{F}$ and $\langle p, r, q \rangle \in \mathbb{F}$. Informally, a favoring relation corresponds to *comparative reasons* (e.g., reasons for you to marry Hugh *rather than* Hugo), so the asymmetry condition corresponds to the claim that nothing can be a reason both for q rather than r and for r rather than q. The favoring relation is used in §5 to define semantic validity.

On a given interpretation m, a declarative sentence p is *true* ($m \vDash p$) or not ($m \nvDash p$), and an imperative sentence i is *satisfied* ($m \Vdash_s i$) or not ($m \nVdash_s i$); if i is not satisfied, then it is either *violated* ($m \Vdash_v i$) or *avoided* ($m \Vdash_a i$). Specifically:

Truth of a declarative sentence on an interpretation

(C1) $m \vDash e$ iff $e \in \mathbb{S}$.

(C2) $m \vDash \ulcorner \sim p \urcorner$ iff $m \nvDash p$.

(C3) $m \vDash \ulcorner p \;\&\; q \urcorner$ iff both $m \vDash p$ and $m \vDash q$.

(C4) $m \vDash \ulcorner p \vee q \urcorner$ iff either $m \vDash p$ or $m \vDash q$ (or both).

(C5) $m \vDash \ulcorner p \rightarrow q \urcorner$ iff either $m \nvDash p$ or $m \vDash q$.

(C6) $m \vDash \ulcorner p \leftrightarrow q \urcorner$ iff either both $m \vDash p$ and $m \vDash q$ or both $m \nvDash p$ and $m \nvDash q$.

Satisfaction, violation, and avoidance of an imperative sentence on an interpretation

(C7) $m \Vdash_s \ulcorner !p \urcorner$ iff $m \vDash p$, and $m \Vdash_v \ulcorner !p \urcorner$ iff $m \nvDash p$.

(C8) $m \Vdash_s \ulcorner \sim i \urcorner$ iff $m \Vdash_v i$, and $m \Vdash_v \ulcorner \sim i \urcorner$ iff $m \Vdash_s i$.

(C9) $m \Vdash_s \ulcorner i \;\&\; j \urcorner$ iff either both $m \Vdash_s i$ and $m \nVdash_v j$ or both $m \Vdash_s j$ and $m \nVdash_v i$, and $m \Vdash_v \ulcorner i \;\&\; j \urcorner$ iff either $m \Vdash_v i$ or $m \Vdash_v j$.

(C10) $m \Vdash_s \ulcorner i \vee j \urcorner$ iff either $m \Vdash_s i$ or $m \Vdash_s j$, and $m \Vdash_v \ulcorner i \vee j \urcorner$ iff either both $m \Vdash_v i$ and $m \nVdash_s j$ or both $m \Vdash_v j$ and $m \nVdash_s i$.

(C11) $m \Vdash_s \ulcorner p \to i \urcorner$ iff both $m \vDash p$ and $m \Vdash_s i$, and $m \Vdash_v \ulcorner p \to i \urcorner$ iff both $m \vDash p$ and $m \Vdash_v i$.

(C12) $m \Vdash_s \ulcorner i \to p \urcorner$ iff both $m \nvDash p$ and $m \Vdash_v i$, and $m \Vdash_v \ulcorner i \to p \urcorner$ iff both $m \nvDash p$ and $m \Vdash_s i$.

(C13) $m \Vdash_s \ulcorner p \leftrightarrow i \urcorner$ iff either both $m \vDash p$ and $m \Vdash_s i$ or both $m \nvDash p$ and $m \Vdash_v i$, and $m \Vdash_v \ulcorner p \leftrightarrow i \urcorner$ iff either both $m \vDash p$ and $m \Vdash_v i$ or both $m \nvDash p$ and $m \Vdash_s i$.

(C14) $m \Vdash_s \ulcorner i \leftrightarrow p \urcorner$ iff $m \Vdash_s \ulcorner p \leftrightarrow i \urcorner$, and $m \Vdash_v \ulcorner i \leftrightarrow p \urcorner$ iff $m \Vdash_v \ulcorner p \leftrightarrow i \urcorner$.

(C15) $m \Vdash_a i$ iff both $m \nVdash_s i$ and $m \nVdash_v i$.

Note that, for any m and i, $m \Vdash_s i$ only if $m \nVdash_v i$. See [8, pp. 532–45] for a detailed defense of C7–C15. A *contradiction* is either a declarative sentence that is false (i.e., not true) on every interpretation or an imperative sentence that is violated on every interpretation. Sentences φ and ψ are *logically equivalent* (i.e., $\varphi \Leftrightarrow \psi$) only if either they are both declarative or they are both imperative. For declarative sentences p and q, $p \Leftrightarrow q$ iff, for any $m, m \vDash p$ iff $m \vDash q$. (Equivalently, $p \Leftrightarrow q$ iff p and q are interderivable in classical sentential logic.) For imperative sentences i and j, $i \Leftrightarrow j$ iff, for any m, both (1) $m \Vdash_s i$ iff $m \Vdash_s j$ and (2) $m \Vdash_v i$ iff $m \Vdash_v j$.

Theorem 3.1 (Semantic Replacement). *For any imperative sentence i and any sentences φ and ψ, if φ is a subsentence of i and $\varphi \Leftrightarrow \psi$, then $i \Leftrightarrow i(\varphi/\psi)$ — where $i(\varphi/\psi)$ is any sentence that results from replacing in i at least one occurrence of φ with ψ.*

Proof. The proof is by induction on the number of occurrences of connectives in i. For the base step, take any i in which no connectives occur. Then, for some e, i is $\ulcorner !e \urcorner$. So if, for some p, $e \Leftrightarrow p$, then $i(e/p)$, namely $\ulcorner !p \urcorner$, is logically equivalent to $\ulcorner !e \urcorner$: for any $m, m \Vdash_s \ulcorner !p \urcorner$ iff $m \vDash p$ iff $m \vDash e$ iff $m \Vdash_s \ulcorner !e \urcorner$ (and similarly $m \Vdash_v \ulcorner !p \urcorner$ iff $m \Vdash_v \ulcorner !e \urcorner$). For the inductive step, take any natural number n and suppose (*induction hypothesis*) that, for any i with at most n occurrences of connectives, and any φ and ψ such that φ is a subsentence of i and $\varphi \Leftrightarrow \psi$, $i \Leftrightarrow i(\varphi/\psi)$. To complete the proof, take any i with at most $n + 1$ occurrences of connectives and any φ and ψ such that φ is a *proper* subsentence of i (the case in which φ is i is trivial) and $\varphi \Leftrightarrow \psi$. To prove that $i \Leftrightarrow i(\varphi/\psi)$, there are eight cases to consider.

Case 1: i is $\ulcorner !p \urcorner$. Then φ is a subsentence of p, and $i(\varphi/\psi)$ is $\ulcorner !p(\varphi/\psi) \urcorner$. By classical sentential logic, $p(\varphi/\psi) \Leftrightarrow p$. It follows, similarly to the base case, that $i \Leftrightarrow i(\varphi/\psi)$.

Case 2: i is $\ulcorner \sim j \urcorner$. Then φ is a subsentence of j, and $i(\varphi/\psi)$ is $\ulcorner \sim j(\varphi/\psi) \urcorner$. By the induction hypothesis, $j \Leftrightarrow j(\varphi/\psi)$ (because j has at most n occurrences of connectives). It follows that $i \Leftrightarrow i(\varphi/\psi)$: for any m, $m \Vdash_s i$ iff $m \Vdash_v j$ iff $m \Vdash_v j(\varphi/\psi)$ iff $m \Vdash_s \ulcorner \sim j(\varphi/\psi) \urcorner$ iff $m \Vdash_s i(\varphi/\psi)$ (and similarly $m \Vdash_v i$ iff $m \Vdash_v i(\varphi/\psi)$).

Case 3: i is $\ulcorner j \ \&\ k \urcorner$. Then φ is a subsentence of j or of k (or both). Suppose it is only of j (if it is only of k, or of both j and k, the proof proceeds similarly). Then $i(\varphi/\psi)$ is $\ulcorner j(\varphi/\psi) \ \&\ k \urcorner$. By the induction hypothesis, $j \Leftrightarrow j(\varphi/\psi)$ (because j has at most n occurrences of connectives). It follows that $i \Leftrightarrow i(\varphi/\psi)$: for any $m, m \Vdash_s i$ iff (either both $m \Vdash_s j$ and $m \nVdash_v k$ or both $m \Vdash_s k$ and $m \nVdash_v j$) iff (either both $m \Vdash_s j(\varphi/\psi)$ and $m \nVdash_v k$ or both $m \Vdash_s k$ and $m \nVdash_v j(\varphi/\psi)$) iff $m \Vdash_s \ulcorner j(\varphi/\psi) \ \&\ k \urcorner$ iff $m \Vdash_s i(\varphi/\psi)$ (and similarly $m \Vdash_v i$ iff $m \Vdash_v i(\varphi/\psi)$).

The proof proceeds similarly in the remaining five cases, namely the cases in which i is $\ulcorner p \vee j \urcorner, \ulcorner p \rightarrow j \urcorner, \ulcorner j \rightarrow p \urcorner, \ulcorner p \leftrightarrow j \urcorner$, or $\ulcorner j \leftrightarrow p \urcorner$, so I omit those cases for the sake of brevity. □

4 Replacement interderivability

In this section, I define replacement derivations, and I prove that there is a replacement derivation of j from i iff $i \Leftrightarrow j$.

Definition 4.1. *For any imperative sentences i and j:*

1. *A* replacement derivation *of j from i is a finite sequence of imperative sentences (called the* lines *of the derivation) such that (a) the last line is j, (b) the first line is i, and (c) each line except the first can be obtained from the previous line by applying once a replacement rule from Table 1.*

2. *i and j are* replacement interderivable *(i.e., $i \dashv\vdash j$) iff there is a replacement derivation of j from i.*

In Table 1, and in what follows, '$p \dashv\vdash_{CSL} q$' abbreviates "p and q are interderivable in classical sentential logic", and for any sentences φ and ψ, '$\varphi \bowtie \psi$' abbreviates "from any imperative sentence k, one can obtain $k(\varphi/\psi)$ if φ is a subsentence of k, and one can obtain $k(\psi/\varphi)$ if ψ is a subsentence of k". For simplicity, I omit corner quotes in tables.

Theorem 4.2 (Soundness of Replacement Rules)**.** *For any imperative sentences i and j, if $i \bowtie j$ according to a replacement rule in Table 1, then $i \Leftrightarrow j$.*

Proof. For the sake of brevity, I examine only EX, ME, and IC; the proof is similar for the other replacement rules.

Exportation: For any m, $m \Vdash_s \ulcorner p \rightarrow (q \rightarrow i) \urcorner$ iff — by C11 — (both $m \vDash p$ and $m \Vdash_s \ulcorner q \rightarrow i \urcorner$) iff ($m \vDash p, m \vDash q$, and $m \Vdash_s i$) iff — by C3 — (both $m \vDash \ulcorner p \ \&\ q \urcorner$ and $m \Vdash_s i$) iff $m \Vdash_s \ulcorner (p \ \&\ q) \rightarrow i \urcorner$ (and similarly $m \Vdash_v \ulcorner p \rightarrow (q \rightarrow i) \urcorner$ iff $m \Vdash_v \ulcorner (p \ \&\ q) \rightarrow i \urcorner$).

Name of rule and abbreviation		Rule
Declarative Replacement	DR	If $p \dashv\vdash_{CSL} q$, then $p \bowtie q$
Transposition	TR	$i \to p \quad \bowtie \quad \sim p \to \sim i$
Negated Conditional	NC	$\sim (p \to i) \quad \bowtie \quad p \to \ \sim i$
Exportation	EX	$p \to (q \to i) \quad \bowtie \quad (p \,\&\, q) \to i$
Commutativity	CO	$p \leftrightarrow i \quad \bowtie \quad i \leftrightarrow p$
Material Equivalence	ME	$p \leftrightarrow i \quad \bowtie \quad (p \to i) \,\&\, (i \to p)$
Absorption	AB	$p \to !q \quad \bowtie \quad p \to !(p \,\&\, q)$
Tautologous Antecedent	TA	$(p \vee \sim p) \to i \quad \bowtie \quad i$
Unconditional Negation	UN	$\sim !p \quad \bowtie \quad ! \sim p$
Imperative Conjunction	IC	$(p \to !q) \,\&\, (p' \to !q') \quad \bowtie \quad (p \vee p') \to !((p \to q) \,\&\, (p' \to q'))$
Imperative Disjunction	ID	$(p \to !q) \vee (p' \to !q') \quad \bowtie \quad (p \vee p') \to !((p \,\&\, q) \vee (p' \,\&\, q'))$

Table 1: Replacement rules

Material Equivalence: Note first that, (1) if $m \Vdash_s \ulcorner p \to i \urcorner$ (i.e., both $m \vDash p$ and $m \Vdash_s i$), then $m \nVdash_v \ulcorner i \to p \urcorner$ (i.e., it is not the case that both $m \nvDash p$ and $m \Vdash_s i$). Similarly, (2) if $m \Vdash_s \ulcorner i \to p \urcorner$, then $m \nVdash_v \ulcorner p \to i \urcorner$. Now, for any m: $m \Vdash_s \ulcorner p \leftrightarrow i \urcorner$ iff — by C13 — (either both $m \vDash p$ and $m \Vdash_s i$ or both $m \nvDash p$ and $m \Vdash_v i$) iff — by C11 and C12 — (either $m \Vdash_s \ulcorner p \to i \urcorner$ or $m \Vdash_s \ulcorner i \to p \urcorner$) iff — by (1) and (2) — (either both $m \Vdash_s \ulcorner p \to i \urcorner$ and $m \nVdash_v \ulcorner i \to p \urcorner$ or both $m \Vdash_s \ulcorner i \to p \urcorner$ and $m \nVdash_v \ulcorner p \to i \urcorner$) iff — by C9 — $m \Vdash_s \ulcorner (p \to i) \,\&\, (i \to p) \urcorner$ (and similarly $m \Vdash_v \ulcorner p \leftrightarrow i \urcorner$ iff $m \Vdash_v \ulcorner (p \to i) \,\&\, (i \to p) \urcorner$).

Imperative Conjunction: Note first that $m \Vdash_s \ulcorner p \to !q \urcorner$ iff (both $m \vDash p$ and $m \Vdash_s !q$) iff (both $m \vDash p$ and $m \vDash q$) iff $m \vDash \ulcorner p \,\&\, q \urcorner$. Similarly, $m \Vdash_v \ulcorner p \to !q \urcorner$ iff $m \vDash \ulcorner p \,\&\, \sim q \urcorner$. Now, for any m: $m \Vdash_s \ulcorner (p \to !q) \,\&\, (p' \to !q') \urcorner$ iff — by C9 — (either both $m \Vdash_s \ulcorner p \to !q \urcorner$ and $m \nVdash_v \ulcorner p' \to !q' \urcorner$ or both $m \Vdash_s \ulcorner p' \to !q' \urcorner$ and $m \nVdash_v \ulcorner p \to !q \urcorner$) iff (either both $m \vDash \ulcorner p \,\&\, q \urcorner$ and $m \nvDash \ulcorner p' \,\&\, \sim q' \urcorner$ or both $m \vDash \ulcorner p' \,\&\, q' \urcorner$ and $m \nvDash \ulcorner p \,\&\, \sim q \urcorner$) iff $m \vDash \ulcorner ((p \,\&\, q) \,\&\, \sim (p' \,\&\, \sim q')) \vee ((p' \,\&\, q') \,\&\, \sim (p \,\&\, \sim q)) \urcorner$ iff — by classical sentential logic — $m \vDash \ulcorner (p \vee p') \,\&\, ((p \to q) \,\&\, (p' \to q')) \urcorner$ iff $m \Vdash_s \ulcorner (p \vee p') \to !((p \to q) \,\&\, (p' \to q')) \urcorner$ (and similarly for violation). □

Theorem 4.3 (Syntactic Replacement)**.** *For any imperative sentences i, j, and k, if j is a subsentence of i and $j \dashv\vdash k$, then $i \dashv\vdash i(j/k)$.*

Proof. Suppose $j \dashv\vdash k$. The proof is by induction on the number of lines of a replacement derivation. For the base step, suppose there is a one-line replacement derivation of k from j. Then j is the same sentence as k and thus $i \dashv\vdash i(j/k)$. For the inductive step, take any non-zero natural number n and suppose (*induction hypothesis*) that, if there is a replacement derivation with n lines of k from j, then $i \dashv\vdash i(j/k)$. To complete the proof, take any replacement derivation with $n + 1$ lines of k from j. Then k can be obtained from the n-th line k' by applying once a replacement rule, so k is $k'(\varphi/\psi)$, where φ is a subsentence of k' and ψ is a sentence such that $\varphi \bowtie \psi$. Let i' be the sentence that results from replacing with k' in i exactly those occurrences of j that are replaced with k in i to get $i(j/k)$. By the induction hypothesis, (1) $i \dashv\vdash i'$. Since k is $k'(\varphi/\psi)$, $i(j/k)$ results from replacing in i' some occurrences of φ with ψ. So $i(j/k)$ is $i'(\varphi/\psi)$, and thus — since $\varphi \bowtie \psi$ — (2) $i(j/k)$ can be obtained from i' by applying once a replacement rule. By (1) and (2), $i \dashv\vdash i(j/k)$. □

Theorem 4.4 (Canonical Form)**.** *For any imperative sentence i, there are declarative sentences p and q such that $i \dashv\vdash \ulcorner p \rightarrow\, !q \urcorner$.*

Proof. The proof is by induction on the number of occurrences of connectives in i. For the base step, take any i in which no connectives occur. Then, for some e, i is $\ulcorner !e \urcorner$, and then, by TA (see Table 1), $i \dashv\vdash \ulcorner (e \vee \sim e) \rightarrow\, !e \urcorner$. For the inductive step, take any natural number n and suppose (*induction hypothesis*) that, for any i with at most n occurrences of connectives, there are p and q such that $i \dashv\vdash \ulcorner p \rightarrow\, !q \urcorner$. To complete the proof, take any i with at most $n + 1$ occurrences of connectives. There are eight cases to consider.

Case 1: i is $\ulcorner !p \urcorner$. Then, by TA, $i \dashv\vdash \ulcorner (p \vee \sim p) \rightarrow\, !p \urcorner$.

Case 2: i is $\ulcorner \sim j \urcorner$. Then j has at most n occurrences of connectives and thus, by the induction hypothesis, $j \dashv\vdash \ulcorner p \rightarrow\, !q \urcorner$ (for some p and q; I omit such remarks in what follows). Then, by Theorem 4.3, $i \dashv\vdash \ulcorner \sim (p \rightarrow\, !q) \urcorner$, and then, by NC and UN, $i \dashv\vdash \ulcorner p \rightarrow\, ! \sim q \urcorner$.

Case 3: i is $\ulcorner j \,\&\, k \urcorner$. Then j has at most n occurrences of connectives and thus, by the induction hypothesis, $j \dashv\vdash \ulcorner p \rightarrow\, !q \urcorner$. Similarly, $k \dashv\vdash \ulcorner p' \rightarrow\, !q' \urcorner$. So, by Theorem 4.3, $i \dashv\vdash \ulcorner (p \rightarrow\, !q) \,\&\, (p' \rightarrow\, !q') \urcorner$, and thus, by IC, $i \dashv\vdash \ulcorner (p \vee p') \rightarrow\, !((p \rightarrow q) \,\&\, (p' \rightarrow q')) \urcorner$.

Case 4: i is $\ulcorner j \vee k \urcorner$. Then, similarly to case 3, $i \dashv\vdash \ulcorner (p \rightarrow\, !q) \vee (p' \rightarrow\, !q') \urcorner$, and thus, by ID, $i \dashv\vdash \ulcorner (p \vee p') \rightarrow\, !((p \,\&\, q) \vee (p' \,\&\, q')) \urcorner$.

Case 5: i is $\ulcorner p \rightarrow j \urcorner$. Then, by the induction hypothesis, $j \dashv\vdash \ulcorner q \rightarrow\, !r \urcorner$. So, by Theorem 4.3, $i \dashv\vdash \ulcorner p \rightarrow (q \rightarrow\, !r) \urcorner$, and thus, by EX, $i \dashv\vdash \ulcorner (p \,\&\, q) \rightarrow\, !r \urcorner$.

Case 6: i is $\ulcorner j \rightarrow p \urcorner$. Then, similarly to case 5, $i \dashv\vdash \ulcorner (q \rightarrow\, !r) \rightarrow p \urcorner$, and thus, by TR, NC, EX, and UN, $i \dashv\vdash \ulcorner (\sim p \,\&\, q) \rightarrow\, ! \sim r \urcorner$.

Case 7: i is $\ulcorner p \leftrightarrow j \urcorner$. Then, similarly to case 5, $i \dashv\vdash \ulcorner p \leftrightarrow (q \rightarrow !r) \urcorner$, and thus, by ME, $i \dashv\vdash \ulcorner (p \rightarrow (q \rightarrow !r))\ \&\ ((q \rightarrow !r) \rightarrow p) \urcorner$. So, by the replacement rules used in case 6, $i \dashv\vdash \ulcorner ((p\ \&\ q) \rightarrow !r)\ \&\ ((\sim p\ \&\ q) \rightarrow !\sim r) \urcorner$, and thus, by IC, $i \dashv\vdash \ulcorner ((p\ \&\ q) \vee (\sim p\ \&\ q)) \rightarrow !(((p\ \&\ q) \rightarrow r)\ \&\ ((\sim p\ \&\ q) \rightarrow\ \ \sim r)) \urcorner$.

Case 8: i is $\ulcorner j \leftrightarrow p \urcorner$. Then, by CO, $i \dashv\vdash \ulcorner p \leftrightarrow j \urcorner$, and the proof proceeds as in case 7. □

Theorem 4.5 (Soundness and Completeness for Replacement Interderivability)**.** *For any imperative sentences i and j, $i \Leftrightarrow j$ if (soundness) and only if (completeness) $i \dashv\vdash j$.*

Proof. **Proof of Soundness.** Suppose $i \dashv\vdash j$. The proof is by induction on the number of lines of a replacement derivation. For the base step, suppose there is a one-line replacement derivation of j from i. Then i is the same sentence as j and thus $i \Leftrightarrow j$. For the inductive step, take any non-zero natural number n and suppose (*induction hypothesis*) that, if there is a replacement derivation with n lines of j from i, then $i \Leftrightarrow j$. To complete the proof, take any replacement derivation with $n + 1$ lines of j from i. Then j can be obtained from the n-th line k by applying once a replacement rule, so j is $k(\varphi/\psi)$, where φ is a subsentence of k and ψ is a sentence such that $\varphi \bowtie \psi$. By the induction hypothesis, (1) $i \Leftrightarrow k$. By Theorem 4.2, $\varphi \Leftrightarrow \psi$ if φ and ψ are imperative sentences; if they are declarative, then j can be obtained from k by applying once DR, so $\varphi \dashv\vdash_{CSL} \psi$ and thus again $\varphi \Leftrightarrow \psi$. By Theorem 3.1, $k \Leftrightarrow k(\varphi/\psi)$; i.e., (2) $k \Leftrightarrow j$. By (1), (2), and the transitivity of logical equivalence (which follows from its definition in §3), $i \Leftrightarrow j$.

Proof of Completeness. Suppose $i \Leftrightarrow j$. By Theorem 4.4, there are p, q, p', and q' such that (1) $i \dashv\vdash \ulcorner p \rightarrow !q \urcorner$ and thus (by soundness) $i \Leftrightarrow \ulcorner p \rightarrow !q \urcorner$, and (2) $j \dashv\vdash \ulcorner p' \rightarrow !q' \urcorner$ and thus $j \Leftrightarrow \ulcorner p' \rightarrow !q' \urcorner$. Then (3) $\ulcorner p \rightarrow !q \urcorner \Leftrightarrow \ulcorner p' \rightarrow !q' \urcorner$. It follows that $p \Leftrightarrow p'$: for any $m, m \vDash p$ iff (either both $m \vDash p$ and $m \vDash q$ or both $m \vDash p$ and $m \nvDash q$) iff — by C11 — (either $m \Vdash_s \ulcorner p \rightarrow !q \urcorner$ or $m \Vdash_v \ulcorner p \rightarrow !q \urcorner$) iff — by (3) — (either $m \Vdash_s \ulcorner p' \rightarrow !q' \urcorner$ or $m \Vdash_v \ulcorner p' \rightarrow !q' \urcorner$) iff (either both $m \vDash p'$ and $m \vDash q'$ or both $m \vDash p'$ and $m \nvDash q'$) iff $m \vDash p'$. Since $p \Leftrightarrow p'$, (4) $p \dashv\vdash_{CSL} p'$. One can show similarly that $\ulcorner p\ \&\ q \urcorner \Leftrightarrow \ulcorner p'\ \&\ q' \urcorner$, so (5) $\ulcorner p\ \&\ q \urcorner \dashv\vdash_{CSL} \ulcorner p'\ \&\ q' \urcorner$. To conclude: i is replacement interderivable, by (1), with $\ulcorner p \rightarrow !q \urcorner$, and thus also, by AB, with $\ulcorner p \rightarrow !(p\ \&\ q) \urcorner$, and thus also, by (4) and DR, with $\ulcorner p' \rightarrow !(p\ \&\ q) \urcorner$, and thus also, by (5) and DR, with $\ulcorner p' \rightarrow !(p'\ \&\ q') \urcorner$, and thus also, by AB, with $\ulcorner p' \rightarrow !q' \urcorner$, and thus finally, by (2), with j. □

Corollary 4.6 (of Theorems 4.4 and 4.5)**.** *For any imperative sentence i, there are declarative sentences p and q such that $i \Leftrightarrow \ulcorner p \rightarrow !q \urcorner$.*

Corollary 4.7 (of Theorems 4.4 and 4.5)**.** *For any imperative sentence i, there are declarative sentences s and v such that, for any m, $m \vDash s$ iff $m \Vdash_s i$ and $m \vDash v$ iff $m \Vdash_v i$.*

Proof. By Corollary 4.6, there are p and q such that $i \Leftrightarrow \ulcorner p \rightarrow !q \urcorner$. Then, for any $m, m \Vdash_s i$ iff $m \Vdash_s \ulcorner p \rightarrow !q \urcorner$ iff (by C11, C7, and C3) $m \vDash \ulcorner p \;\&\; q \urcorner$, so take s to be $\ulcorner p \;\&\; q \urcorner$. Similarly, take v to be $\ulcorner p \;\&\; \sim q \urcorner$. □

5 Strong and weak semantic validity

A *pure imperative argument* (of the language of SPIL) is an ordered pair $\langle \Gamma, i \rangle$, where Γ is a non-empty finite set of imperative sentences (the *premises* of the argument) and i is an imperative sentence (the *conclusion* of the argument). In this paper, I do not examine arguments whose premises and conclusions include both declarative and imperative sentences (e.g., the argument $\langle \{A \rightarrow !B, A\}, \; !B \rangle$). Building on previous work [10, 12], I say that (roughly) a pure imperative argument is semantically valid when, on every interpretation, its conclusion is "supported" by everything that supports its premises. Also building on previous work, I distinguish *strong* from *weak* support — and, correspondingly, strong from weak semantic validity — as follows:

Definition 5.1. *For any declarative sentence p, any imperative sentence i, and any interpretation m:*

1. *p* strongly supports *i on m iff (a) $m \vDash p$, (b) i is not a contradiction, and (c) $\langle p, q, r \rangle \in \mathbb{F}$ for any q and r that are not both contradictions and are such that, for any m', both (i) $m' \vDash q$ only if $m' \Vdash_s i$ and (ii) $m' \vDash r$ only if $m' \Vdash_v i$.*

2. *p* weakly supports *i on m iff p strongly supports on m some j such that, for any m', both (a) $m' \Vdash_s j$ only if $m' \Vdash_s i$ and (b) $m' \Vdash_a i$ iff $m' \Vdash_a j$.*

Definition 5.2. *A pure imperative argument $\langle \Gamma, i \rangle$ is (1)* strongly semantically valid *(i.e., $\Gamma \Vdash_s i$) iff, for any m, every p that* strongly *supports on m every conjunction*[2] *of all members of* Γ *also* strongly *supports i on m, and is (2)* weakly semantically valid *(i.e., $\Gamma \Vdash_w i$) iff, for any m, every p that* weakly *supports on m every conjunction of all members of* Γ *also* weakly *supports i on m.*

It follows from Definition 5.1 that, if p strongly supports i on m, then p also weakly supports i on m. Informally, the distinction between strong and weak semantic validity

[2] See [10, pp. 396–8] for an explanation of why I define semantic validity in terms of supporting *conjunctions* of all premises and not in terms of supporting *every* premise. Given the intensionality condition (§3) and the logical equivalence of any two conjunctions of all premises of an argument, supporting (strongly or weakly, on an interpretation) *some* conjunction of all premises of an argument amounts to supporting *every* conjunction of all premises of the argument. Because sentences are *finite* strings of symbols, I do not define conjunctions of infinitely many sentences (contrast [12, p. 1706, n. 1]; this is why I defined an argument as having finitely many premises.

captures a conflict of intuitions about whether, for example, “sign the letter” entails “sign or burn the letter”: one can show that the pure imperative argument $\langle\{!S\}, !(S \vee B)\rangle$ is weakly but not strongly semantically valid.[3]

Theorem 5.3 (Semantic Equivalence)**.** *For any imperative sentences i and j:*

1. *$i \Vdash_s j$ (i.e., $\{i\} \Vdash_s j$) iff either i is a contradiction or, for any m, both (a) $m \Vdash_s j$ only if $m \Vdash_s i$ and (b) $m \Vdash_v j$ only if $m \Vdash_v i$;*

2. *$i \Vdash_w j$ iff, for any m, both (a) $m \Vdash_a i$ only if $m \Vdash_a j$ and (b) $m \Vdash_v j$ only if $m \Vdash_v i$.*

Proof. The theorem provides necessary and sufficient conditions for strong and for weak semantic validity. The proof has four parts, and is similar to the proof in Appendix A of [10].

First part: Sufficient condition for strong semantic validity. If i is a contradiction, then (by Definition 5.1) no p strongly supports i on any m, and then (by Definition 5.2) $i \Vdash_s j$. If, for any m', both (a) $m' \Vdash_s j$ only if $m' \Vdash_s i$ and (b) $m' \Vdash_v j$ only if $m' \Vdash_v i$, take any $m = \langle \mathbb{S}, \mathbb{F} \rangle$ and any p. Suppose that (1) p strongly supports i on m. Then (2) $m \vDash p$ (by Definition 5.1) and (3) j is not a contradiction (because, by Definition 5.1, i is not a contradiction; so, for some $m', m' \nVdash_v i$, and thus — by (b) — $m' \nVdash_v j$). Moreover, (4) for any q and r, if q and r are not both contradictions and are such that, for any m', both (i) $m' \vDash q$ only if $m' \Vdash_s j$ and (ii) $m' \vDash r$ only if $m' \Vdash_v j$, then $\langle p, q, r \rangle \in \mathbb{F}$ (by (1) and Definition 5.1, because (by (i) and (a)) $m' \vDash q$ only if $m' \Vdash_s i$ and (by (ii) and (b)) $m' \vDash r$ only if $m' \Vdash_v i$). By (2), (3), (4), and Definition 5.1, p strongly supports j on m, so (by Definition 5.2) $i \Vdash_s j$.

Second part: Necessary condition for strong semantic validity. By Corollary 4.7, there are declarative sentences s and v such that, for any m, $m \vDash s$ iff $m \Vdash_s i$ and $m \vDash v$ iff $m \Vdash_v i$, and declarative sentences s' and v' that satisfy the corresponding conditions with respect to j. Suppose, for reductio, that (1) $i \Vdash_s j$ but (2) i is not a contradiction and (3) it is not the case that, for every m, both (a) $m \Vdash_s j$ only if $m \Vdash_s i$ and (b) $m \Vdash_v j$ only if $m \Vdash_v i$.

[3] Defending the above definitions lies beyond the scope of this paper: I have extensively defended in previous work [10, 12] an account of validity on which the definitions are based. I say that the definitions are “based” on my previously defended account of validity because that account is about “arguments” whose premises and conclusions are not sentences of a formal language, but are instead what imperative and declarative sentences of natural languages typically express, namely *prescriptions* (i.e., commands, requests, instructions, suggestions, etc.) and *propositions* respectively. Deviating slightly from previous work in order to keep my definition of an interpretation (§3) simple, I formulated Definition 5.1 so that it has as consequences two claims corresponding to what in previous work I understood as assumptions about favoring, namely the claims that (1) no declarative sentence strongly supports on any interpretation an imperative sentence which is a contradiction (cf. Assumption 1 in [10, p. 433]) and (2) every declarative sentence that is true on an interpretation strongly supports on that interpretation any *semantically empty* imperative sentence (cf. [12, p. 1708, n. 6]), namely any imperative sentence that is avoided on every interpretation.

Consider an interpretation $m = \langle \mathbb{S}, \mathbb{F} \rangle$, where $\mathbb{S} = \{e\}$ for some e (so (4) $m \vDash e$) and $\mathbb{F}$ is the set of ordered triples $\langle p, q, r \rangle$ such that (i) $p \Leftrightarrow e$, (ii) $q \vdash_{CSL} s$ (i.e., s is derivable from q in classical sentential logic; equivalently, for any $m', m' \vDash q$ only if $m' \Vdash_s i$), (iii) $r \vdash_{CSL} v$ (equivalently, for any $m', m' \vDash r$ only if $m' \Vdash_v i$), and (iv) q and r are not both contradictions. $\mathbb{F}$ satisfies the asymmetry condition (§3): if one supposes for reductio that both $\langle p, q, r \rangle \in \mathbb{F}$ and $\langle p, r, q \rangle \in \mathbb{F}$, then one gets that q and r are both contradictions (contradicting (iv)): q is a contradiction because, for any m', if $m' \vDash q$, then both $m' \Vdash_s i$ and $m' \Vdash_v i$ (which is impossible), and similarly for r. The intensionality condition (§3) is also satisfied. By (2), (4), the definition of $\mathbb{F}$, and Definition 5.1, e strongly supports i on m. Then, by (1) and Definition 5.2, (5) e also strongly supports j on m. Let q be $\ulcorner s' \ \& \ \sim s \urcorner$ and r be $\ulcorner v' \ \& \ \sim v \urcorner$. By (3), q and r are not both contradictions. Moreover, for any m', $m' \vDash q$ only if $m' \Vdash_s j$, and $m' \vDash r$ only if $m' \Vdash_v j$. Then, by (5) and Definition 5.1, $\langle e, q, r \rangle \in \mathbb{F}$. By (ii), $\ulcorner s' \ \& \ \sim s \urcorner \vdash_{CSL} s$, so (there is no interpretation on which $\ulcorner (s' \ \& \ \sim s) \ \& \ \sim s \urcorner$ is true, and thus) $s' \vdash_{CSL} s$; equivalently, (6) for any m, $m \Vdash_s j$ only if $m \Vdash_s i$. Similarly, by (iii), $\ulcorner v' \ \& \ \sim v \urcorner \vdash_{CSL} v$, so $v' \vdash_{CSL} v$; equivalently, (7) for any m, $m \Vdash_v j$ only if $m \Vdash_v i$. But (6) and (7) together contradict (3), and the reductio is complete.

Third part: Sufficient condition for weak semantic validity. Suppose that, for any m, both (a) $m \Vdash_a i$ only if $m \Vdash_a j$ and (b) $m \Vdash_v j$ only if $m \Vdash_v i$. Take any m and any p that weakly supports i on m. By Definition 5.1, (1) p strongly supports on m some imperative sentence i^* such that, for any m', both (i) $m' \Vdash_s i^*$ only if $m' \Vdash_s i$ and (ii) $m' \Vdash_a i$ iff $m' \Vdash_a i^*$. Let k be $\ulcorner (s' \vee v') \rightarrow !(s^* \ \& \ s') \urcorner$, where s' and v' are as in the second part of the proof and s^* is a declarative sentence such that, for any m', $m' \vDash s^*$ iff $m' \Vdash_s i^*$ (see Corollary 4.7). Then, (2) for any m', $m' \Vdash_s k$ only if $m' \vDash \ulcorner s^* \ \& \ s' \urcorner$, and thus, (3) for any $m', m' \Vdash_s k$ only if $m' \Vdash_s i^*$. Moreover, (4) for any m', $m' \Vdash_v k$ only if $m' \Vdash_v i^*$ (as one can show by using (a), (b), (i), and (ii); see [10, p. 436, n. 68]). By (3), (4), and the first part of the proof, (5) $i^* \Vdash_s k$. By (1), (5), and Definition 5.2, (6) p strongly supports k on m. But, (7) for any m', $m' \Vdash_s k$ only if $m' \Vdash_s j$ (by (2)), and, (8) for any m', $m' \Vdash_a j$ iff $m' \Vdash_a k$ (because $m' \Vdash_a k$ iff $m' \nvDash \ulcorner s' \vee v' \urcorner$). By (6), (7), (8), and Definition 5.1, p weakly suports j on m, so (by Definition 5.2) $i \Vdash_w j$.

Fourth part: Necessary condition for weak semantic validity. Suppose, for reductio, that (1) $i \Vdash_w j$ but (2) either (a) for some m, both $m \Vdash_a i$ and $m \nVdash_a j$, or (b) for some m, both $m \Vdash_v j$ and $m \nVdash_v i$ (i.e., it is not the case that, for every m, both (a') $m \Vdash_a i$ only if $m \Vdash_a j$ and (b') $m \Vdash_v j$ only if $m \Vdash_v i$). By (2), i is not a contradiction (i.e., for some $m, m \nVdash_v i$; this is immediate if (b) is true, and follows from $m \Vdash_a i$ if (a) is true). Consider an interpretation $m = \langle \mathbb{S}, \mathbb{F} \rangle$ defined as in the second part of the proof. As in that part, e strongly supports i on m, so e also weakly supports i on m. Then, by (1) and Definition 5.2, e also weakly supports j on m. Then, by Definition 5.1, (3) e strongly supports on m some i^* such that, for any m', both (i) $m' \Vdash_s i^*$ only if $m' \Vdash_s j$ and (ii) $m' \Vdash_a j$ iff $m' \Vdash_a i^*$. By (2), for some m, $m \nVdash_a j$ (this is immediate if (a) is true, and follows from $m \Vdash_v j$ if (b) is true). Then, by (ii), for some $m, m \nVdash_a i^*$, so s^* and v^* are not both contradictions — where s^* is as in the third part of the proof and v^* is a declarative sentence such that, for any m, $m \vDash v^*$ iff $m \Vdash_v i^*$ (see Corollary 4.7). Then, by (3) and Definition 5.1, $\langle e, s^*, v^* \rangle \in \mathbb{F}$, and by the definition of $\mathbb{F}$ in the second part of the proof, (4) $s^* \vdash_{CSL} s$ and (5) $v^* \vdash_{CSL} v$. But then (a) is false: for any m, if $m \nVdash_a j$ and thus (by (ii)) $m \nVdash_a i^*$, then $m \nVdash_a i$ (because either $m \Vdash_s i^*$, and then by (4) $m \Vdash_s i$ and thus $m \nVdash_a i$, or $m \Vdash_v i^*$, and then by (5) $m \Vdash_v i$ and thus $m \nVdash_a i$). Moreover, (b) is false: for any m, if $m \Vdash_v j$ (and thus (6) $m \nVdash_s j$ and (7) $m \nVdash_a j$), then $m \Vdash_v i$ (because $m \nVdash_s i^*$, by (i) and (6), and $m \nVdash_a i^*$, by (ii) and (7), so $m \Vdash_v i^*$ and, by (5), $m \Vdash_v i$). The falsity of (a) and (b) contradicts (2), and the reductio is complete. □

Corollary 5.4 (of Theorem 5.3)**.** *For any imperative sentences i and j, (1) i* $\Vdash_s j$ *only if i* $\Vdash_w j$*, and (2) i* $\Leftrightarrow$ *j iff (both i* $\Vdash_s j$ *and j* $\Vdash_s i$*) iff (both i* $\Vdash_w j$ *and j* $\Vdash_w i$*).*

6 Strong and weak derivability

In this section, I define strong and weak derivations, and I prove that there is a strong (or weak) derivation of i from Γ iff the argument $\langle \Gamma, i \rangle$ is strongly (or weakly) semantically valid.

Definition 6.1. *For any pure imperative argument* $\langle \Gamma, i \rangle$*:*

1. *A* strong derivation *of i from* Γ *is a finite sequence of imperative sentences (called the* lines *of the derivation) such that (a) the last line is i and (b) each line either is a conjunction of* all *members of* Γ *or can be obtained from a previous line by applying once either a replacement rule from Table 1 or a pure imperative inference rule (other than ICE) from Table 2.*

2. *A* weak derivation *of i from* Γ *is a finite sequence of imperative sentences (called the* lines *of the derivation) such that (a) the last line is i and (b) each line either is (a member or) a conjunction of members of* Γ *or can be obtained from a previous line by applying once either a replacement rule from Table 1 or a pure imperative inference rule from Table 2.*

3. $\langle \Gamma, i \rangle$ *is (a)* strongly syntactically valid *(i.e.,* $\Gamma \vdash_s i$*) iff there is a strong derivation of i from* Γ*, and is (b)* weakly syntactically valid *(i.e.,* $\Gamma \vdash_w i$*) iff there is a weak derivation of i from* Γ*.*

In Table 2, and in what follows, for any imperative sentences i and j, '$i \triangleright j$' abbreviates "from i, one can obtain j". It follows from Definition 6.1 that every strong derivation is a weak derivation, so $\Gamma \vdash_s i$ only if $\Gamma \vdash_w i$. Moreover, since replacement rules may be applied in strong derivations, $j \dashv\vdash i$ only if $j \vdash_s i$ (i.e., $\{j\} \vdash_s i$). Note two differences between

Name of rule and abbreviation		**Rule**
Ex Contradictione Quodlibet	ECQ	$!(p \,\&\, \sim p) \triangleright i$
Declarative Antecedent Introduction	DAI	$i \triangleright p \rightarrow i$
Imperative Conjunction Elimination	ICE	$i \,\&\, j \triangleright i$

Table 2: Pure imperative inference rules

weak and strong derivations. First, all pure imperative inference rules in Table 2 may be applied in a weak derivation, but Imperative Conjunction Elimination (ICE) may *not* be applied in a strong derivation. The motivation behind this difference is that, for example, the argument $\langle\{!A \,\&\, !B\}, !A\rangle$ is (weakly but) not strongly semantically valid (as one can show by using Theorem 5.3), but strong derivations are intended to correspond to strong semantic validity. Second, any single premise can be the first line of a weak derivation, but no single premise (as opposed to a conjunction of all premises) can be the first line of a strong derivation (unless there is only one premise). The motivation behind this difference is that, for example, the argument $\langle\{!A, !B\}, !A\rangle$ is (weakly but) not strongly semantically valid (see [10, p. 397]).[4]

Theorem 6.2 (Soundness of Inference Rules)**.** *For any declarative sentence p and any imperative sentences i and j: (1)* $\ulcorner !(p \,\&\, \sim p)\urcorner \Vdash_s i$*; (2)* $i \Vdash_s \ulcorner p \rightarrow i\urcorner$*; (3)* $\ulcorner i \,\&\, j\urcorner \Vdash_w i$*.*

Proof. (1) Since $\ulcorner !(p \,\&\, \sim p)\urcorner$ is a contradiction, $\ulcorner !(p \,\&\, \sim p)\urcorner \Vdash_s i$ by Theorem 5.3. (2) For any m, both (a) $m \Vdash_s \ulcorner p \rightarrow i\urcorner$ only if $m \Vdash_s i$ (by C11) and (b) $m \Vdash_v \ulcorner p \rightarrow i\urcorner$ only if $m \Vdash_v i$ (by C11), so $i \Vdash_s \ulcorner p \rightarrow i\urcorner$ by Theorem 5.3. (3) For any m, both (a) $m \Vdash_a \ulcorner i \,\&\, j\urcorner$ only if $m \Vdash_a i$ (by C9 and C15) and (b) $m \Vdash_v i$ only if $m \Vdash_v \ulcorner i \,\&\, j\urcorner$ (by C9), so $\ulcorner i \,\&\, j\urcorner \Vdash_w i$ by Theorem 5.3. □

Theorem 6.3 (Strengthening the Antecedent and Weakening the Consequent)**.** *For any declarative sentences* p, p', q*, and* q'*, and any imperative sentence i: (1) if* $p' \vdash_{CSL} p$*, then* $\ulcorner p \rightarrow i\urcorner \vdash_s \ulcorner p' \rightarrow i\urcorner$*; (2) if* $q \vdash_{CSL} q'$*, then* $\ulcorner p \rightarrow !q\urcorner \vdash_w \ulcorner p \rightarrow !q'\urcorner$*.*

Proof. (1) $\ulcorner p \rightarrow i\urcorner \vdash_s \ulcorner p' \rightarrow (p \rightarrow i)\urcorner$ (by DAI), and $\ulcorner p' \rightarrow (p \rightarrow i)\urcorner \vdash_s \ulcorner (p' \,\&\, p) \rightarrow i\urcorner$ (by EX). But if $p' \vdash_{CSL} p$, then $\ulcorner p' \,\&\, p\urcorner \dashv\vdash_{CSL} p'$, and then $\ulcorner (p' \,\&\, p) \rightarrow i\urcorner \vdash_s \ulcorner p' \rightarrow i\urcorner$

[4] DAI is redundant given ICE, AB, IC, and EX. Indeed, $\ulcorner p \rightarrow i\urcorner$ can be obtained by ICE from $\ulcorner (p \rightarrow i) \,\&\, (\sim p \rightarrow i)\urcorner$, which is replacement interderivable with i: i is replacement interderivable with $\ulcorner q \rightarrow !r\urcorner$ (for some q and r, by Theorem 4.4), and thus also with $\ulcorner q \rightarrow !(q \,\&\, r)\urcorner$ (by AB), and thus also with $\ulcorner q \rightarrow !(q \,\&\, (q \rightarrow r))\urcorner$ (by DR, since $\ulcorner q \,\&\, r\urcorner \dashv\vdash_{CSL} \ulcorner q \,\&\, (q \rightarrow r)\urcorner$), and thus also with $\ulcorner q \rightarrow !(q \rightarrow r)\urcorner$ (by AB), and thus also with $\ulcorner ((p \,\&\, q) \vee (\sim p \,\&\, q)) \rightarrow !(((p \,\&\, q) \rightarrow r) \,\&\, ((\sim p \,\&\, q) \rightarrow r))\urcorner$ (by DR, since $q \dashv\vdash_{CSL} \ulcorner (p \,\&\, q) \vee (\sim p \,\&\, q)\urcorner$ and $\ulcorner q \rightarrow r\urcorner \dashv\vdash_{CSL} \ulcorner ((p \,\&\, q) \rightarrow r) \,\&\, ((\sim p \,\&\, q) \rightarrow r)\urcorner$), and thus also with $\ulcorner ((p \,\&\, q) \rightarrow !r) \,\&\, ((\sim p \,\&\, q) \rightarrow !r)\urcorner$ (by IC), and thus also with $\ulcorner (p \rightarrow (q \rightarrow !r)) \,\&\, (\sim p \rightarrow (q \rightarrow !r))\urcorner$ (by EX), and thus finally with $\ulcorner (p \rightarrow i) \,\&\, (\sim p \rightarrow i)\urcorner$ (by Theorem 4.3). It does not follow, however, that DAI is redundant in *strong* derivations: ICE may *not* be applied in strong derivations.

(by DR). (2) If $q \vdash_{CSL} q'$, then $\ulcorner q' \,\&\, q \urcorner \dashv\vdash_{CSL} q$. Then there is a weak derivation from $\ulcorner p \rightarrow\, !q \urcorner$ of $\ulcorner p \rightarrow\, !(q' \,\&\, q) \urcorner$ (by DR), and thus also of $\ulcorner p \rightarrow\, !(p \,\&\, (q' \,\&\, q)) \urcorner$ (by AB), and thus also of $\ulcorner (p \vee p) \rightarrow\, !((p \vee p) \,\&\, ((p \rightarrow q') \,\&\, (p \rightarrow q))) \urcorner$ (by DR, since $p \dashv\vdash_{CSL} \ulcorner p \vee p \urcorner$ and $\ulcorner p \,\&\, (q' \,\&\, q) \urcorner \dashv\vdash_{CSL} \ulcorner (p \vee p) \,\&\, ((p \rightarrow q') \,\&\, (p \rightarrow q)) \urcorner$), and thus also of $\ulcorner (p \vee p) \rightarrow\, !((p \rightarrow q') \,\&\, (p \rightarrow q)) \urcorner$ (by AB), and thus also of $\ulcorner (p \rightarrow\, !q') \,\&\, (p \rightarrow\, !q) \urcorner$ (by IC), and thus finally of $\ulcorner p \rightarrow\, !q' \urcorner$ (by ICE). □

Theorem 6.4 (Soundness and Completeness for Strong and Weak Derivability)**.** *For any pure imperative argument* $\langle \Gamma, i \rangle$*, (1)* $\Gamma \Vdash_s i$ *if (soundness) and only if (completeness)* $\Gamma \vdash_s i$*, and (2)* $\Gamma \Vdash_w i$ *if (soundness) and only if (completeness)* $\Gamma \vdash_w i$*.*

Proof. **Proof of Soundness**. The proof is by induction on the number of lines of a strong or weak derivation. For the base step, suppose there is a one-line strong (case 1) or weak (case 2) derivation of i from Γ. In case 1, i is a conjunction of all members of Γ and thus (by Definition 5.2) $\Gamma \Vdash_s i$. In case 2, i is (a member or) a conjunction of members of Γ; so, if i is not a conjunction of *all* members of Γ (if it is, the proof proceeds as in case 1), there is a conjunction j of the remaining members of Γ, and $\ulcorner i \,\&\, j \urcorner$ is a conjunction of all members of Γ. Then $\Gamma \Vdash_w i$ because, by Definition 5.2, $\Gamma \Vdash_w \ulcorner i \,\&\, j \urcorner$, and by Theorem 6.2, $\ulcorner i \,\&\, j \urcorner \Vdash_w i$. For the inductive step, take any non-zero natural number n and suppose (*induction hypothesis*) that: (case 1) if there is a strong derivation with at most n lines of i from Γ, then $\Gamma \Vdash_s i$; (case 2) if there is a weak derivation with at most n lines of i from Γ, then $\Gamma \Vdash_w i$. To complete the proof, take any strong (case 1) or weak (case 2) derivation with at most $n + 1$ lines of i from Γ. Suppose that i is *not* a conjunction of all (case 1) or some (case 2) members of Γ (if it is, the proof proceeds as in the base step). Then i can be obtained from an n'-th line j ($n' \leq n$) by applying once (case 1) ECQ, DAI, or a replacement rule, or (case 2) any inference or replacement rule. Then (1_s) $j \Vdash_s i$ in case 1 (by Theorem 6.2) and (1_w) $j \Vdash_w i$ case 2 (by Theorem 6.2 and Corollary 5.4). By the induction hypothesis and the fact that the sequence of the first n' lines of the strong (case 1) or weak (case 2) derivation of i from Γ is a strong (case 1) or weak (case 2) derivation with at most n lines of j from Γ, (2_s) $\Gamma \Vdash_s j$ in case 1, and (2_w) $\Gamma \Vdash_w j$ in case 2. By (1_s), (2_s), and the transitivity of strong semantic validity (which follows from Definition 5.2), $\Gamma \Vdash_s i$ in case 1. Similarly, by (1_w), (2_w), and the transitivity of weak semantic validity, $\Gamma \Vdash_w i$ in case 2.

Proof of Completeness. Take any pure imperative argument $\langle \Gamma, i \rangle$ and any conjunction i' of all members of Γ. By Theorem 4.4, there are p, q, p', and q' such that (1) $i \dashv\Vdash \ulcorner p \rightarrow\, !q \urcorner$ and (2) $i' \dashv\Vdash \ulcorner p' \rightarrow\, !q' \urcorner$. By (1), (2), and Theorem 4.5: (3) for any m, $m \Vdash_s i$ iff $m \vDash \ulcorner p \,\&\, q \urcorner$, $m \Vdash_s i'$ iff $m \vDash \ulcorner p' \,\&\, q' \urcorner$, $m \Vdash_v i$ iff $m \vDash \ulcorner p \,\&\, \sim q \urcorner$, and $m \Vdash_v i'$ iff $m \vDash \ulcorner p' \,\&\, \sim q' \urcorner$ (see the proof of Corollary 4.7). *Case 1*: $\Gamma \Vdash_s i$. Then (4) $i' \Vdash_s i$ (by Definition 5.2). Case 1a: i' is a contradiction. Then, for any r, $i' \Leftrightarrow \ulcorner !(r \,\&\, \sim r) \urcorner$ (since i' and $\ulcorner !(r \,\&\, \sim r) \urcorner$ are both violated on every m) and thus (by Theorem 4.5) $i' \dashv\Vdash$

$\ulcorner !(r \,\&\, \sim r)\urcorner$, so $i' \vdash_s \ulcorner !(r \,\&\, \sim r)\urcorner$. Then there is a strong derivation of i from i' (and thus from Γ), since i can be obtained from $\ulcorner !(r \,\&\, \sim r)\urcorner$ by ECQ. Case 1b: i' is not a contradiction. Then, by (4) and Theorem 5.3: (5) for any m, $m \Vdash_s i$ only if $m \Vdash_s i'$, and (6) for any m, $m \Vdash_v i$ only if $m \Vdash_v i'$. By (3) and (5): (7) $\ulcorner p \,\&\, q\urcorner \vdash_{CSL} \ulcorner p' \,\&\, q'\urcorner$. By (3) and (6): (8) $\ulcorner p \,\&\, \sim q\urcorner \vdash_{CSL} \ulcorner p' \,\&\, \sim q'\urcorner$. By using (7), (8), and classical sentential logic, one can show that (9) $p \vdash_{CSL} p'$ and (10) $\ulcorner p \,\&\, (p' \,\&\, q')\urcorner \dashv\vdash_{CSL} \ulcorner p \,\&\, q\urcorner$. To conclude: there is a strong derivation from Γ of i' (by Definition 6.1), and thus also of $\ulcorner p' \rightarrow !q'\urcorner$ (by (2)), and thus also of $\ulcorner p' \rightarrow !(p' \,\&\, q')\urcorner$ (by AB), and thus also of $\ulcorner p \rightarrow !(p' \,\&\, q')\urcorner$ (by (9) and Theorem 6.3), and thus also of $\ulcorner p \rightarrow !(p \,\&\, (p' \,\&\, q'))\urcorner$ (by AB), and thus also of $\ulcorner p \rightarrow !(p \,\&\, q)\urcorner$ (by (10) and DR), and thus also of $\ulcorner p \rightarrow !q\urcorner$ (by AB), and thus finally of i (by (1)). *Case 2*: $\Gamma \Vdash_w i$. Then $i' \Vdash_w i$ (by Definition 6.1 and the observation that any member or conjunction of members of Γ can be obtained from i' by applying replacement rules or ICE or both). Then, by Theorem 5.3: (11) for any m, $m \Vdash_a i'$ only if $m \Vdash_a i$, and (12) for any m, $m \Vdash_v i$ only if $m \Vdash_v i'$. By (3) and (11): (13) $p \vdash_{CSL} p'$. By (3) and (12): (14) $\ulcorner p \,\&\, \sim q\urcorner \vdash_{CSL} \ulcorner p' \,\&\, \sim q'\urcorner$. By (14) and classical sentential logic: (15) $\ulcorner p \,\&\, q'\urcorner \vdash_{CSL} q$. To conclude: there is a weak derivation from Γ of i' (by Definition 6.1), and thus also of $\ulcorner p' \rightarrow !q'\urcorner$ (by (2)), and thus also of $\ulcorner p \rightarrow !q'\urcorner$ (by (13) and Theorem 6.3), and thus also of $\ulcorner p \rightarrow !(p \,\&\, q')\urcorner$ (by AB), and thus also of $\ulcorner p \rightarrow !q\urcorner$ (by (15) and Theorem 6.3), and thus finally of i (by (1)).[5] □

7 Conclusion

I conclude by noting that in future work I plan to address some of the limitations of SPIL by presenting sound and complete natural deduction systems for three further logics: (1) *First-Order Pure Imperative Logic* (FOPIL), which includes quantifiers and identity but no modal operators; (2) *Sentential Modal Imperative Logic* (SMIL), which includes modal operators but no quantifiers or identity and deals with arguments from declarative or imper-

[5] Hansen [7] provides an alternative sound and complete natural deduction system for SPIL. More precisely, Hansen considers a language of SPIL in which every imperative sentence is either of the form $\ulcorner !q\urcorner$ or of the form $\ulcorner p \rightarrow !q\urcorner$ (Hansen uses '⇒' instead of '→'). This limitation is not crucial: by Theorem 4.4, every imperative sentence of the language of SPIL is inderderivable with a sentence of the form $\ulcorner p \rightarrow !q\urcorner$ by using only replacement rules (which Hansen does not introduce, although in effect he relies on TA and one of his inference rules corresponds to IC). Hansen's system has six inference rules; five of them correspond to (special cases of) ECQ, IC, Strengthening the Antecedent, and Weakening the Consequent, but the remaining rule is new. (Only the rule that corresponds to a special case of Weakening the Consequent may not be applied in Hansen's "strong deductions", which roughly correspond to strong derivations.) Here is the new rule (which Hansen calls "Contextual Extensionality") in my notation: if $p \vdash_{CSL} \ulcorner q \leftrightarrow r\urcorner$, then $\ulcorner p \rightarrow !q\urcorner \triangleright \ulcorner p \rightarrow !r\urcorner$. Although this rule has no analog in my system, its effects can be simulated by using only replacement rules: if $p \vdash_{CSL} \ulcorner q \leftrightarrow r\urcorner$, then $\ulcorner p \,\&\, q\urcorner \dashv\vdash_{CSL} \ulcorner p \,\&\, r\urcorner$, and then $\ulcorner p \rightarrow !(p \,\&\, q)\urcorner$ and $\ulcorner p \rightarrow !(p \,\&\, r)\urcorner$ are replacement interderivable (by DR), and thus so are also $\ulcorner p \rightarrow !q\urcorner$ and $\ulcorner p \rightarrow !r\urcorner$ (by AB).

ative premises (or both) to declarative or imperative conclusions; and (3) *First-Order Modal Imperative Logic* (FOMIL), which combines (1) and (2).

References

[1] Clarke, David S., Jr. (1973). *Deductive Logic: An Introduction to Evaluation Techniques and Logical Theory*. Carbondale, IL: Southern Illinois University Press.

[2] Clarke, David S., Jr., & Behling, Richard (1998). *Deductive Logic: An Introduction to Evaluation Techniques and Logical Theory* (2nd ed.). Lanham, MD: University Press of America.

[3] Fine, Kit (2018). Compliance and command I—Categorical Imperatives. *The Review of Symbolic Logic, 11,* 609–633.

[4] Gensler, Harry J. (1990). *Symbolic Logic: Classical and Advanced Systems*. Englewood Cliffs, NJ: Prentice-Hall.

[5] Gensler, Harry J. (1996). *Formal Ethics*. New York: Routledge.

[6] Gensler, Harry J. (2002). *Introduction to Logic*. New York: Routledge.

[7] Hansen, Jörg (2014). Be nice! How simple imperatives simplify imperative logic. *Journal of Philosophical Logic, 43,* 965–977.

[8] Vranas, Peter B. M. (2008). New foundations for imperative logic I: Logical connectives, consistency, and quantifiers. *Noûs, 42,* 529-572.

[9] Vranas, Peter B. M. (2010). In defense of imperative inference. *Journal of Philosophical Logic, 39,* 59–71.

[10] Vranas, Peter B. M. (2011). New foundations for imperative logic: Pure imperative inference. *Mind, 120,* 369–446.

[11] Vranas, Peter B. M. (2013). Imperatives, logic of. In H. LaFollette (Ed.), *International Encyclopedia of Ethics* (Vol. 5, pp. 2575–2585). Oxford: Blackwell.

[12] Vranas, Peter B. M. (2016). New foundations for imperative logic III: A general definition of argument validity. *Synthese, 193*, 1703–1753.

Received 13 June 2018

Formal Periodic Steady-State Analysis of Power Converters in Time-Domain

Asad Ahmed
School of Electrical Engineering and Computer Science (SEECS)
National University of Sciences and Technology (NUST), Islamabad, Pakistan
asad.ahmed@seecs.nust.edu.pk

Osman Hasan
School of Electrical Engineering and Computer Science (SEECS)
National University of Sciences and Technology (NUST), Islamabad, Pakistan
osman.hasan@seecs.nust.edu.pk

Ammar Hasan
School of Electrical Engineering and Computer Science (SEECS)
National University of Sciences and Technology (NUST), Islamabad, Pakistan
ammar.hasan@seecs.nust.edu.pk

Abstract

Time-domain based periodic steady-state analysis is an indispensable component to analyze switching functionality and design specifications of power electronics converters. Traditionally, paper-and-pencil proof methods and computer-based tools are used to conduct the time-domain based steady-state analysis of these converters. However, these techniques do not provide an accurate analysis due to their inability to model and analyze continuous behaviors exhibited by the power electronics converters. On the other hand, an accurate analysis is direly needed in many safety and cost-critical power electronics applications, such as biomedical, hybrid electric vehicles, and aerospace engineering. To alleviate the issues pertaining to the above-mentioned techniques, we propose a methodology, based on higher-order-logic theorem proving, to conduct the time-domain based steady-state analysis of power electronics converters in this paper. The proposed methodology is primarily based on a formalized switching function analysis technique, ordinary linear differential equations and steady-state conditions of the systems. To illustrate the usefulness of proposed formalization, we present the formal time-domain steady-state analysis of a commonly used DC-DC Buck converter.

Vol. 6 No. 3 2019
Journal of Applied Logics — IfCoLog Journal of Logics and their Applications

1 Introduction

Power electronics converters are an integral part of, almost, every realizable electrical/electronics system, as a power processing stage, to meet their power requirements [10]. These systems are typically composed of semiconductor devices, like switches, energy storage and dissipative elements, i.e., inductors, capacitors, and resistors, and integrated circuits for control. Generally, periodic steady-state analysis is a mandatory preprocessing step for the small-signal analysis, which is used to evaluate the performance of the converter. Moreover, time-domain based analysis is necessary for the study of the switching functionality, which is central to the power conversion operation of the converters [10]. However, switching is a highly non-linear phenomenon and therefore leads to significant modeling, analysis and design challenges of these systems.

Traditionally, paper-and-pencil proof methods or computer-based numerical techniques are used to perform the time-domain based steady-state analysis of the power electronics systems. The paper-and-pencil proofs are usually based on many assumptions, such as small-ripple approximations, and averaging techniques to linearize the nonlinear behavior of the systems to analyze the systems in steady-state [10]. These linearized models, expressed as ordinary linear differential equations, are then simulated using a variety of computer based simulation tools, such as MATLAB Simulink, Saber, PSpice, to evaluate the performance of the power electronics systems. Generally, these computer based simulation tools use discretized time or frequency domain models of the systems and numerical integration methods [7] for solving the differential equations of the converters [8]. Therefore, the above-mentioned techniques cannot ascertain an accurate and reliable analysis of the power converters due to inherent approximation based nature of these techniques. For example, the accuracy of paper-and-pencil proof methods is usually limited by the underlying approximate linearized model. On the other hand, the nonlinear analysis is, mathematically, not tractable and due to human involvement is highly likely error prone. Similarly, the numerical methods employed in the simulation techniques, based upon the discretization of time or frequency, lead to truncation errors and also cannot accurately model the hybrid behavior, i.e., continuous behavior driven by discrete events, exhibited by power converters [22]. To address this issue, computer algebra systems, which are software programs for the symbolic processing of mathematical expressions, are also employed for the analysis of such systems [16]. However, the symbolic processing is based on the unverified program codes, and therefore prone to bugs [21]. Thus, given the aforementioned inaccuracies, these traditional techniques should not be relied upon for the analysis of power electronics systems, especially when they are used in safety-critical areas, such as implantable medical devices [3] and automotive industry [9], and mission-critical areas, such as aerospace engineering [13], where bugs may lead to heavy monetary or human life loss.

In recent years, formal methods have been extensively employed for the accurate analysis of a variety of hardware and software systems. The transfer function of DC-DC converters has been verified [6] in the frequency domain using higher-order-logic theorem proving based on the signal flow graph and Mason's gain formula. The transfer function is then used to reason about the efficiency, stability and resonance of pulse width modulation push-pull DC-DC converter and 1-boost cell DC-DC converter. However, the nature of formalization does not permit to reason about the interesting features of switch, which is a key element of power electronic converters. Model checking has also been used for the analysis of the DC-DC Buck circuit [18] [20] using a hybrid automaton equivalent model of circuit to verify the reachability and safety properties of the circuit. However, the state-based modeling of the circuit does not allow to describe the exact continuous behavior of power converters circuits. Moreover, the state-space explosion issues also limit the scope of model checking for the verification of continuous and hybrid systems. To the best of our knowledge, there is no formal approach in the literature that explicitly allows us to verify the nonlinear aspects pertaining to the modeling and time-domain based steady-state analysis of power electronics systems.

The main motivation of this paper is to develop a formal logical framework for the time-domain based steady-state analysis of power converters. The main challenge in this direction is to be able to model and analyze the continuous structural or topological changes under the switching action [5], which are usually modeled using the Heaviside step function [1], i.e.,

$$u(t) = \begin{cases} 1 & 0 < t \\ 1/2 & t = 0 \\ 0 & t < 0 \end{cases} \tag{1}$$

The topological changes deter the explicit use of conventional circuit analysis techniques, such as mesh and node analysis, for investigating the implementation of the circuit by using the behavior of its individual components and its overall behavior [17]. Another notable consequence is that the switching action introduces piecewise functions, which are also expressed in terms of the Heaviside step function, in the analysis that in turn cannot be analyzed using linear mathematical techniques based on the Riemann integral theory, such as differential chain rule and integration by part. To tackle the former issue, we propose to use the switching function technique [17], which is a commonly used circuit analysis technique that allows to incorporate the topological changes of the circuit in the analysis. We tackled the piecewise nature of the functions in our formal framework by using the Gauge or Henstock-

Kurzweil integral [15]. The Gauge integral is characterized by the Gauge function for the tagged division of an interval over which the function is to be integrated. This simple, but novel, alteration allows us to integrate the functions with countable singularities or the functions that are continuous but not differentiable everywhere on the given interval. It, particularly, supported us in the formal verification of an interesting notion of the Heaviside step function as a generalized function [14] which is widely used to describe discontinuous phenomena in physics and engineering disciplines. As a generalized function, the Heaviside step function acts as an operator on a test function $f(x)$, which needs to be smooth everywhere, as:

$$\int_a^b h(x-c)f(x) = \int_c^b f(x) \qquad \forall\, a\, b\, c.\ a < c < b \tag{2}$$

The smoothness of test function also plays a pivotal role in the differentiation of the piecewise functions involving the Heaviside step function in the formal time-domain based periodic steady-state analysis of power converters.

Besides these foundations, the proposed formalization is based on the formalizations of linear ordinary differential equations and steady-state conditions. The homogeneous linear differential equations using real analysis have been formalized in HOL to model the cyber-physical systems [19]. In this paper, we have extended the logical framework, presented in [19], to the non-homogeneous linear differential equations using complex analysis to formally model the dynamic behavior of the power converters. We have used the multi-variable integral, differential, transcendental and topological theories to define the steady-state conditions due to the piecewise nature of the functions involved in the analysis.

The formalization in this paper is done using the HOL-Light theorem prover [11], which supports formal reasoning about higher-order logic. The main motivation behind this choice is the availability of reasoning support about multi-variable integral, differential, transcendental and topological theories [12], which are the foremost foundations required for the formalization of time-domain based steady-state analysis of power electronics systems.

The rest of the paper is organized as follows: We describe some preliminaries regarding the periodic steady-state analysis of power electronics converters in Section 2. In Section 3, we present the proposed methodology. The formalization of the switching function technique, ordinary differential equations and steady-state conditions in Section 4. We utilize this formalization to formally verify a Power converter circuit, i.e., DC-DC buck converter in Section 5. Finally, Section 6 concludes the paper.

2 Periodic Steady-state Analysis of Power Converters

Power converter circuits use continuous switching among different circuit configurations to achieve the desired power conversion, such as dc-dc, dc-ac, ac-dc and ac-ac. In each circuit configuration, also called mode or state of the converter, the behavior of the circuit variables can be expressed as differential equations with initial conditions from the previous mode at the switching instance. Therefore, the standard approach for the time-domain analysis of these converters consists of developing the differential equations for each mode of the circuit based on the Kirchoff's voltage or current laws to describe the dynamic behavior of these circuits.

Mathematically, the behavior of these systems can be described as:

$$\begin{aligned} H(t, y_1, y_1^1, ..., y_n^{m_n}) &= p(t) && t \in [t_{n-1}, t_n], n, m_n \in \mathbb{N} \\ y_n^k(t_n) &= y_{n-1}^k(t_{n-1}) && k \in \mathbb{N} \\ y_0^1(t_0) &= 0 \end{aligned} \tag{3}$$

Where, H and p are functions of an independent variable t, a dependent variable y_n and its m_n-th order derivative in the corresponding n-th mode, respectively. In power converters, the time is considered as an independent variable, whereas, the voltage or current of the energy storage components is considered as a dependent variable. The order, i.e., m_n, of an ordinary differential equation of the power converter, in the n-th mode, is determined by the number of energy storage elements constituting the mode. The function $p(t)$ is referred to as a non-homogeneous term, which can be zero or non-zero in the n-th mode, depending upon the presence of source in the n-th mode of a power converter. Initially, the value of dependent variable is considered zero, i.e., $y_0^1(t_0) = 0$, however, later on the value of the dependent variable in one mode becomes an initial value for the next mode, i.e., $y_n^k(t_n) = y_{n-1}^k(t_{n-1})$, when switching instance occurs. Whereas, k is the order of the derivative of the dependent variable evaluated at a specific time instance.

For the brevity of the notion, transient and steady-state time-domain behavior of a DC-DC power converter is presented in Fig. 1, base on the above-mentioned standard approach. DC-DC power converter circuits are designed to step-up or step down the dc voltage levels applied at their input. Fig. 1 shows the output behavior, y_t, of a DC-DC power converter under the switching action represented by a rectangular switch wave form, S_w.

In periodic steady-state, the dependent variables of a power converter circuit attain an equilibrium and repeat the behavior over a time period, T_p, constituting l modes. Mathematically, the periodic steady-state behavior of a power converter

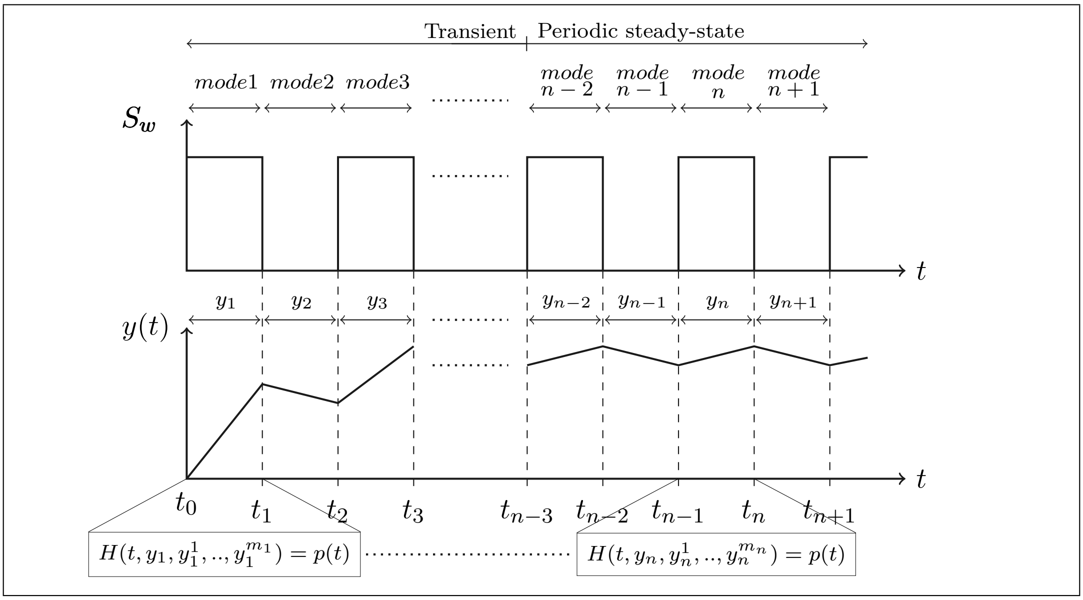


Figure 1: **Dynamic behavior of the output, $y(t)$, of a DC-DC power converter under switching action, represented by the switching wave form, S_w.**

over one time period, when $t \to \infty$, can be represented as:

$$\begin{gathered} H(t, y_n, y_n^1, ..., y_n^{m_n}) = p(t) \quad t \in T, T \in \bigcup_{i=1}^{l} \left[t_{i-1}^{'}, t_i^{'}\right], m_n, n, l \in \mathbb{N} \\ y^k(t_0^{'}) = y^k(t_0^{'} + T_p) \quad T_p = t_{max(i)}^{'} - t_0^{'}, k \in \mathbb{N} \end{gathered} \tag{4}$$

Equation (4) reduces the problem to the identification of the modes in one time period, $T_p = t_{max(i)}^{'} - t_0^{'}$, of the circuit, which is the length of time over which the modes of a power circuit converter repeat themselves. The function y is a piecewise function defined over l modes. Whereas, $y^k(t_0^{'}) = y^k(t_0^{'} + T_p)$ refers to the steady-state conditions of the system variable at reference switching time instances, $t_0^{'}$, and T_p, and k represents the k-th order derivative of the variable.

Fig. 2 illustrates the behavior of the output of a DC-DC power converter in steady-state, which is mathematically modeled in Equation 4. The output, $y(t)$, of the converter exhibits a repetitive behavior over the time period T_p in l modes. In literature, waveforms of the dependent variable, y, are used for the periodic steady-sate analysis of the power converters by applying the principle of inductor volt-second or capacitor-charge, along, with small-ripple approximations to reduce

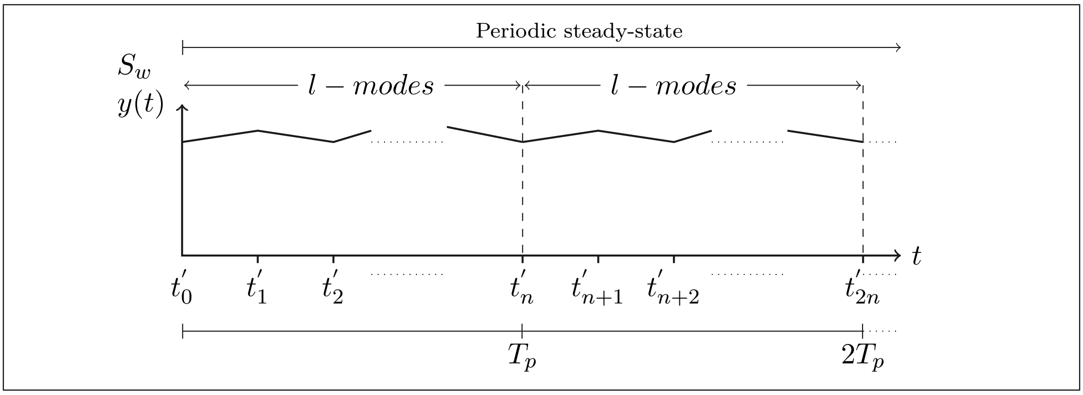


Figure 2: **Behavior of the output, $y(t)$, of a DC-DC power converter in Periodic steady-state.**

the complexity of the analysis by compromising the accuracy [10].

In this paper, we propose a logical framework for the formal verification of the periodic steady-state analysis of power converters in time domain, which are mathematically represented by Equation 4. The challenges to develop a logical framework for the formal verification of the aforementioned problem are two fold. Firstly, we intend to develop a higher-order logic formalization capable of incorporating the topological structural changes over the time period, i.e., $T \in \bigcup_{i=1}^{l} \left[t_{i-1}', t_i'\right]$, thus, enabling us to formally model and reason about the implementation behavior of these circuits within the sound core of the HOL-Light theorem prover. Second we want to develop a formal library of foundations, including; differential equations, concepts from operational calculus described by Equation 2, to formally reason and verify the highly nonlinear behavior of the circuit variables involved in the formal periodic steady-state analysis of these circuits, in higher-order logic. The respective subsections of Section 4 address these challenges by presenting the formalization of switching function technique, differential equations and solution of these differential equations, respectively, to conduct the formal periodic steady-state analysis of power converters in the time-domain.

In the next section, we present the proposed methodology for the formal periodic steady-state analysis of the power converters, in a higher-order-logic theorem prover, i.e., HOL-Light.

3 Proposed Methodology

We propose to use higher-order-logic theorem proving, as shown in Fig. 3, in order to formally verify the power converters operating in the periodic steady-state. The first step in the proposed methodology is to build a formal model for the switching function technique and linear order differential equations to formally express the implementation and specification of power converter circuits, in higher-order logic. The proposed formal modeling of switching function technique is based on the formal definitions of an ideal semiconductor switch, energy storage and dissipative elements, and Kirchoff's current and voltage laws. Whereas, the formal modeling of the linear ordinary differential equation is used for the formal specification of the behavior of each mode of the power converter circuit. The aforementioned two formal models can then be used to formally assert and analyze the implementation of the circuits, as a theorem, using the sound core of HOL-Light. Moreover, the formal specification of ordinary linear differential equations is also used to formally

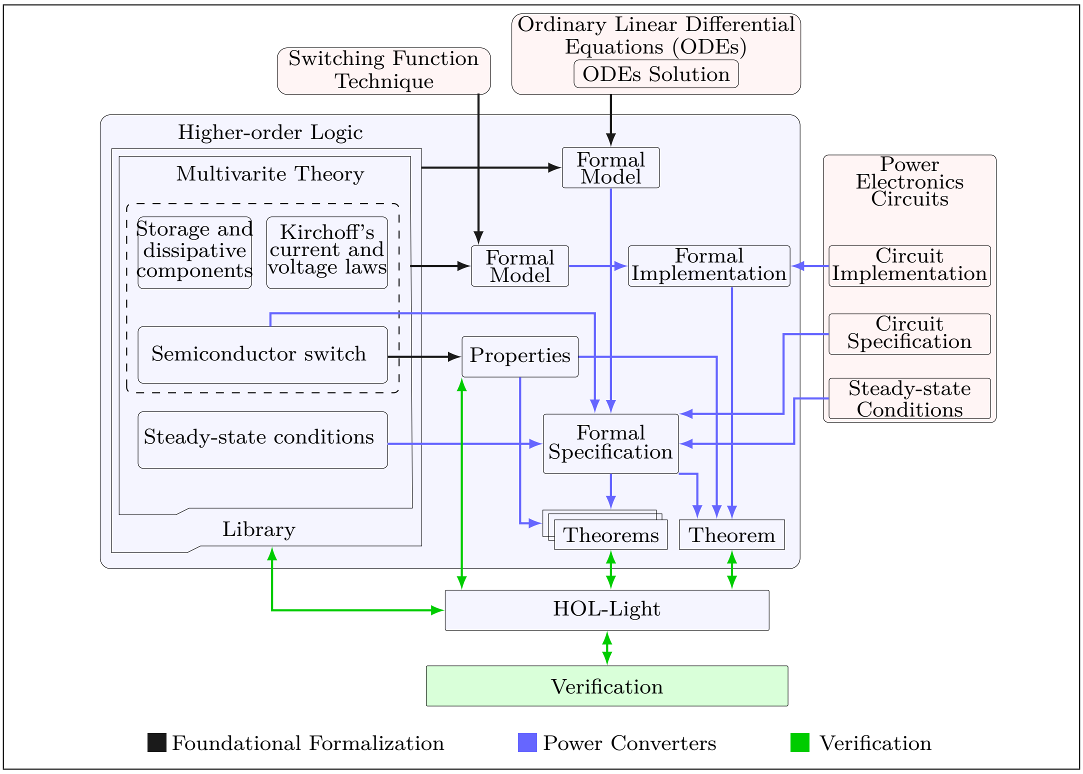


Figure 3: **Proposed Methodology**

verify the correctness of the solutions of these equations. As the steady-state analysis is based upon the formal modeling of the linear ordinary differential equations and their solutions, therefore, in the next step, we propose to formally define the steady-state conditions to conduct the formal analysis of power converters, as shown in Fig 3. These formal definitions, along with multi-variable theories of HOL-Light, are used to formally verify the theorems that are required to conduct the formal steady-state analysis of power converters. Finally, the switch is formalized using the Heaviside step function, and its related properties, such as integration and derivation of piecewise functions involving Heaviside step function, are formally verified. As the switching functionality plays the most vital role in characterizing the nonlinear behavior of the power converters therefore these formally verified properties are used in, almost, every aspect of the formalization and verification.

4 Foundational Formalizations

4.1 Formal Model of the Switching Function Technique

In power converter circuits, semiconductor devices such as, diodes, BJTs (bipolar junction transistors), MOSFETs (metal oxide semiconductor field effect transistors), IGBTs (insulated gate bipolar transistors) etc, are used for performing the switching operation. These semiconductor devices play a vital role in the development of reliable, cost-effective and highly efficient converters [4]. Although, these devices differ in their physics and physical properties, however, as a switch, their function is to connect or disconnect a path or subcircuit, in a converter circuit, to achieve the desired conversion. Therefore, the functionality of an ideal semiconductor device as a switch can be modeled using the Heaviside function, i.e., Equation (1), in HOL-Light:

Definition 1: ⊢ ∀ t. `semi_switch t = if t < &0 then &0 else`
```
                                          (if t = &0 then &1 / &2 else &1)
```

Definition 1 models the functionality of a semiconductor switch as a real value `1`, for connected status, and `0`, for disconnected status, in higher-order logic. Whereas, at the switching instance `t`, it has value `1/2`. The `&` is a typecasting operator in HOL-Light that maps a number to a real number. In our formalization, we use switch status or switching function to refer connected or disconnected switch.

The switching operation is central to the power converters functionality, however, it hinders the straightforward usage of the conventional circuit theory techniques, such as Kirchoff's voltage and current laws. The switching function technique relies on the superposition theorem of the voltage or current to express the behavior

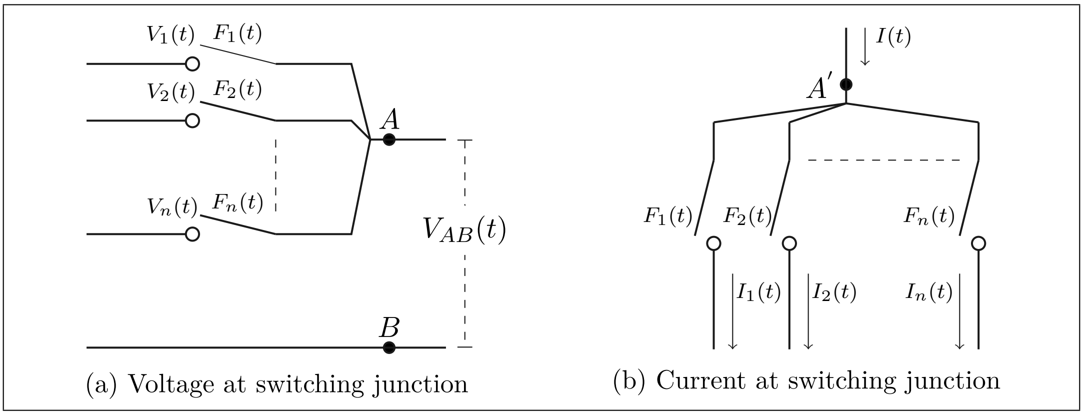


Figure 4: **Switching function technique**

of these quantities in the presence of a switch in the circuit. It is based on the conceptualization of the switch as a modulating function for the input and output power. Based on this notion, the voltages and the currents in the presence of a switch component can be expressed as [1];

$$V_{AB}(t) = \sum_{i=1}^{n} V_i(t)F_i(t) \qquad n \in \mathbb{N} \tag{5a}$$

$$I_i(t) = I(t)\sum_{i=1}^{n} F_i(t) \qquad n \in \mathbb{N} \tag{5b}$$

Equation 5(a), describes voltage at the switch junction, in a mesh, in terms of switching functions. Fig. 4(a) is a pictorial representation of the concept, where n voltage sources are connected to a point, A, through n switches. The voltage, V_{AB}, is then the superposition of the input voltages, however, the contribution of each voltage is dependent upon the associated switching function. Similarly, Equation 5(b), describes the current at a node, $A^{'}$, which has n switches. Fig. 4(b) describes the situation where current, $I(t)$, is supplied to n paths of the circuit through n switches. Each path receives the fraction of total current depending upon its switch status, $F_n(t)$.

Voltages and currents at the switching junction in higher-order logic are defined, as:

Definition 2: ⊢ ∀ mod_lst volt_lst t.
```
switch_volt mod_lst volt_lst t =
vsum (0..LENGTH mod_lst - 1) (λ n.  EL n volt_lst t * Cx (EL n mod_lst))
```

The function `switch_volt` describes the voltage at the switch junction using Equation

5(a). It accepts a list, `volt_lst`, which contains all the possible voltage drops at the switching junction, a list of modes, `mod_lst`, which contains the switch status or switching function for each mode, and `t` is the time, which indicates that this function is time dependent. Whereas, `Cx` is a HOL-Light function, which is used to map a real number, representing the switching function, to a complex number.

Definition 3: ⊢ ∀ `mod_lst curr t.  switch_current mod_lst curr t =`

```
curr t * vsum (0..LENGTH mod_lst - 1) (λ n.  Cx (EL n mod_lst))
```

Definition 3 formally models the current at the switching junction using Equation 5(b). It accepts an argument `curr`, which represents the total supplied current to the switch junction, a list of modes, `mod_lst`, which contains the switch status or switching function for each mode, and `t`, which represents time.

To accomplish the formal modeling of the switching function technique, we also formalize the Kirchoff's voltage and current laws:

Definition 4: ⊢ ∀ `vol_lst t.  kvl vol_lst t =`

```
vsum (0..LENGTH vol_lst - 1) (λn.  EL n vol_lst t) = Cx (&0)
```

Definition 5: ⊢ ∀ `cur_lst t.  kcl cur_lst t =`

```
vsum (0..LENGTH cur_lst - 1) (λn.  EL n cur_lst t) = Cx (&0)
```

The `kvl` and `kcl` functions accept lists of type $(\mathbb{R} \rightarrow \mathbb{C})$, to express the behavior of the time dependent voltages and currents in the given power converter circuit and a time variable `t`. They return the predicates that guarantee that the sum of the voltages in a loop or sum of the currents at a node are zero for all the time instants.

The voltages and currents in Definitions 2 and 3 are piecewise functions due to switching action. We formally verified the result of Equation (2) to conduct the formal analysis involving such functions:

Theorem 1: ⊢ ∀ `f a b c x.`

`A1:(∀t.  (λx.  f (x)) differentiable_on s) ∧`

`A2:∼(real_interval [a,b] = {}) ∧`

`A3:c ∈ [a, b]`

⇒ $\int_a^b$ `(λx.  semi_switch x c ) * f (x)) =` $\int_c^b$ `(λx.  f (x))`

The Assumption `A1` ensures the differentiability of a test function, `f`, over `s`. Whereas, `s`:$(\mathbb{R} \rightarrow \mathbb{B})$ is a set-theoretic definition of the intervals in higher-order logic, over real numbers. For a given real interval `[a,b]`, it represents all possible real intervals, which are subsets of the given real interval. Therefore, Assumption `A1` ensures the differentiability of a test function over all subsets of the given real interval `[a,b]`. Assumptions `A2` and `A3` ensure that the interval is non-empty and point `c` lies within the interval [a, b]. The conclusion of the Theorem 1 formally verifies the affect of applying the Heaviside step function on a test function, i.e., changes the limit of integral. Theorem 1 is formally verified using the formal definition of Gauge integral and its properties, available in HOL-Light theorem prover. This formally verified result plays a very key role in the formal reasoning of the systems which exhibit nonlinear behavior, such as power converters circuits.

The above formalization enables us to formally model and analyze the nonlinear behavior exhibited by the power converters, due to switching action, in higher-order logic.

4.2 Ordinary Linear Differential Equation

An nth-order ordinary linear differential equation can be represented as:

$$a_n(t)\frac{d^n y(t)}{dx} + a_{n-1}(t)\frac{d^{n-1} y(t)}{dx} + ... + a_0(t)y(t) = p(t) \tag{6}$$

We formalized the nth-order derivative function in higher-order logic as follows:

Definition 6:
```
⊢ ∀ n f t. (n_vec_deri 0 f t = f t )  ∧
(∀ n. n_vec_deri (SUC n) f t =
        n_vec_deri n (λ t. vector_derivative f at t) t)
```

The function n_vec_deri accepts a positive integer n that represents the order of the derivative, the function f:(ℝ → ℂ) that represents the complex-valued function that needs to be differentiated, and the variable t:(ℝ) that is the variable with respect to which we want to differentiate the function f. It returns the nth-order derivative of f with respect to t. Now, based on this definition, we can formalize the left-hand side (LHS) and right-hand side (RHS) of Equation (6) in HOL-Light as the following definitions:

Definition 7:
```
⊢ ∀ P y t. diff_eq_lhs A f t =
    vsum (0..LENGTH A) (λ n. Cx ( EL n A t) * n_vec_deri n f t)
```

Definition 8:
```
⊢ ∀ L y t. diff_eq_rhs L p t =
    vsum (0..LENGTH L) (λ n. Cx (EL n L) * EL n p t)
```

In the above definitions, A and L are the coefficient's lists, f:(ℝ → ℂ) and p(t):(ℝ → ℂ) are complex-valued functions, and t:(ℝ) is the time variable to formally model the linear ordinary differential equation. Definition 6 is also used to formally define the steady-state condition of the power converters as:

Definition 9:
```
⊢ ∀ n.  ( steady_state 0 f T_p =
      (n_vec_deri 0 f (&0) = n_vec_deri 0 f T_p) ) ∧
      ( steady_state (SUC n) f T_p =
      (n_vec_deri (SUC n) f (&0) = n_vec_deri (SUC n) f T_p) )
```

The above generic formalization allows to formally model the dynamic behavior of systems represented by differential equations. We have utilized this formalization to formally specify and reason the periodic steady-state behavior of power converters, described in Equation 4.

4.3 Solution of Linear Differential Equations

The general solution to non-homogeneous Equation (6) is expressed as

$$y(t) = y_h(t) + y_p(t) = \sum_{i=1}^{n} c_i y_i(t) + y_p(t) \tag{7}$$

Where, $y_h(t)$ is the linear combination of the fundamental solutions of Equation (6) when $p(t) = 0$, and y_p is the particular solution corresponding to Equation (6) when $p(t) \neq 0$.

The formal verification of the correctness of the solution of linear differential equation, i.e., Equation (6), is based on the linearity property of the derivatives, which we have formally verified for the complex-valued functions as:

Theorem 2:
```
⊢ ∀ n f h t.
  A1:  (λ m t.  m ≤ n ⇒ (λ t.  n_vec_deri m f t) differentiable at t) ∧
  A2:  (λ m t.  m ≤ n ⇒ (λ.  n_vec_deri m h t) differentiable at t)
       ⇒ n_vec_deri n (λt.  Cx a * f t + Cx b * h t) t =
          Cx a * n_vec_deri (λt.  f t) t + Cx b * n_vec_deri (λt.  g t) t
```

We formally verified the solution of a linear differential equation, represented by Equation (7), in the HOL-Light theorem prover as follows:

Theorem 3:
```
⊢ ∀ Y_h C Y_p A L p t.
  A1:  (n_differentiable_fn Y_h (LENGTH A)) ∧
  A2:  (n_differentiable_fn Y_p (LENGTH L)) ∧
  A3:  (n_homo_soln A Y_h t) ∧
  A4:  (n_nonhomo_soln A L Y_h Y_p t)
     ⇒ diff_eq_lhs A (λ t. linear_sol C Y_h t + Y_p t = diff_equ_rhs L p t
```

In Theorem 3, Assumptions `A1` and `A2` ensure the n$^{\text{th}}$-order differentiability of the fundamental solutions, given as a list `Yh`, and particular solution, provided as a list `Yp`, respectively. The predicate in the Assumption `A3`, i.e., `n_order_homo_eq_soln_list`, ensures that each element of the list `Yh` is a solution of the given differential equation, when $p(t) = 0$ in Equation (6), where `L` is the list of coefficients. Similarly, the predicate in Assumption `A4`, i.e., `n_order_nonhomo_eq_soln_list`, ensures that the particular solution, `Yp`, satisfies the differential Equation (6). The function `linear_sol`, used in the conclusion of Theorem 2, models the linear solution combination of fundamental solutions, i.e., $\sum_{i=1}^{n} c_i y_i(t)$, using the lists of solution functions `Yh` and arbitrary constants `C`. The formal verification of Theorem 3 is based on Theorem 1 and the formally verified lemma about solution of homogeneous differential equation, i.e., when $p(t) = 0$ in Equation (6). More details about the modeling and verification steps can be found in our proof script [2]. The formalization, presented in this section, is generic and provides sufficient support to formally model and reason about different aspects of a power converters' circuits including; implementation and behavior, specification, correctness of the solution of differential equations representing the behavior of circuits, and also the steady-state behavior of quantities of interests, such as voltages and currents. The corresponding proof script, which is available for download at [2], has 3000 lines of HOL-Light code and requires about 350 man hours of development time.

5 DC-DC Buck Converter

The DC-DC buck converter is a commonly used power converter that steps down a given input to a desired output level. In a DC-DC Buck converter, operating in a continuous conduction mode, a switch controls the flow of energy from the raw source, Vs, to the output by periodically switching between Positions 1 and 2, as shown in Fig 5. The energy is stored in the inductor when the switch is at Position 1, and is dissipated to the output circuitry, when the switch is at Position 2. The circuit has two modes, i.e., $n = 2$, defined by the switching instances, t_0, t_{on}, and t_{off}. In periodic steady-state the circuit will repeat its behavior periodically over

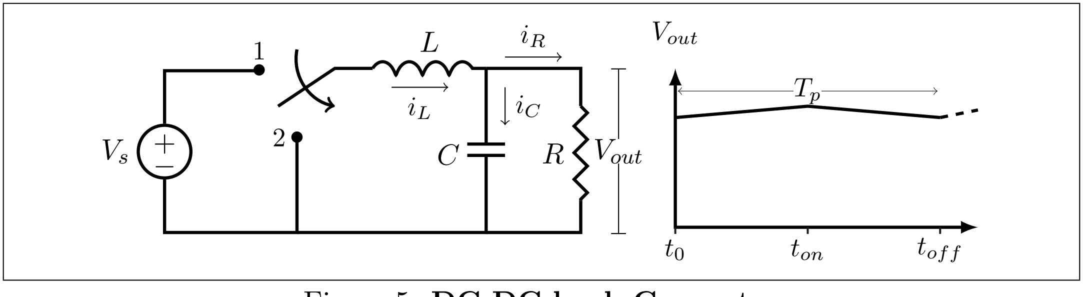


Figure 5: **DC-DC buck Converter**

the time period T_p. Moreover, due to periodic steady-state the dependence on t_0 can be dropped and therefore have assigned $t_0 = 0$ in our analysis. Applying Kirchoff's current and voltage laws in switch Positions 1 and 2, gives the following differential equations for the respective modes:

$$\begin{aligned} i_L &= i_C + i_R \\ \frac{d^2}{dt^2}V^1_{out}(t) + \frac{1}{RC}\frac{d}{dt}V^1_{out}(t) + \frac{1}{LC}V^1_{out}(t) &= \frac{V_s}{LC} \\ V^1_{out}(t) &= c_1 e^{s_1 t} + c_2 e^{s_2 t} + V_s \end{aligned} \tag{8}$$

$$\begin{aligned} i_L &= -i_c - i_R \\ \frac{d^2}{dt^2}V^2_{out}(t) + \frac{1}{RC}\frac{d}{dt}V^2_{out}(t) + \frac{1}{LC}V^2_{out}(t) &= 0 \\ V^2_{out}(t) &= c_3 e^{s_3 t} + c_4 e^{s_4 t} \end{aligned} \tag{9}$$

Where, V_{out} is the output voltage of the converter, as shown in the Fig. 5, and s_1, s_2, s_3 and s_4 are the roots of the characteristic equation of the converter in two modes. Moreover, $s_1 = s_3$ and $s_2 = s_4$ due to the identical characteristic equations. The solution of Equations (8-9), over the time period T_c, can be written using the Heaviside step function as

$$V_{out}(t) = u(t - t_{on})V^1_{out}(t) + (1 - u(t - t_{on}))V^2_{out}(t) \tag{10}$$

In the periodic steady-state, the voltage of the DC-DC buck converter satisfies the following conditions

$$V_{out}(0) = V_{out}(T)\,,\ \frac{d}{dt}V_{out}(0) \quad = \frac{d}{dt}V_{out}(T) \tag{11}$$

The steady-state conditions provide two algebraic equations, however, there are four constants involved in the solution. Two more algebraic equations can be obtained from the continuity of the voltage, i.e., V_{out}, due to continuous conduction mode of the circuit, i.e.,

$$V^1_{out}(t_{on}) = V^2_{out}(t_{on})\,,\ \frac{d}{dt}V^1_{out}(t_{on}) \quad = \frac{d}{dt}V^2_{out}(t_{on}) \tag{12}$$

Equations (11-12) are used to specify the periodic steady-state voltage that allows finding the minimum and peak conduction currents in steady-state. These currents can then be used to determine ripple currents, which are essentially crucial in specifying the components in the design of the converters.

The first step, in the formalization of the DC-DC Buck converter consists of using the switching function technique to write the switch junction voltages, which

Component	Current Relationship
Resistor	$\mathtt{I_R(t)} = \frac{\mathtt{V(t)}}{\mathtt{R}}$
Capacitor	$\mathtt{I_C(t)} = \mathtt{C}\frac{\mathtt{dV(t)}}{\mathtt{d(t)}}$
Inductor	$\mathtt{I_L} = \mathtt{i_0} + \frac{1}{\mathtt{R}}\int_0^t \mathtt{V(t)}$

Table 1: **Basic quantities in DC-DC converter**

in turn requires to formally define the currents of inductor, capacitor and resistor elements. The mathematical expressions for these elements are presented in Table 1, which are formally defined as,

Definition 10: ⊢ ∀ i_o L v. ind_curr v L i_o =
(λ t. i_o + Cx (&1 / L) * integral (interval [&0, t]) v)

Definition 11: ⊢ ∀ C v. cap_curr C v =
(λ t. Cx C * vector_derivative v (at t))

Definition 12: ⊢ ∀ v R. res_curr R v = (λ t. v t * Cx (&1 / R))

Where, `R` , `C` and `L` represent the resistance, capacitance and inductances of the resistor, capacitor and inductor of the circuit. i_o is the initial value of the inductor current, whereas, `v` represents the voltage drop across the circuit elements, at any time `t`. Now, using Definitions 2, 4, 5, 10, 11, and 12, we can formalize the implementation of DC-DC Buck converter as:

Definition 13: ⊢ ∀ i_o L C R V_s V_{out} V_L t_{on} t.
buck_ckt_impl i_o L C R Vs V_{out} V_L t_{on} t =
(Vl = switch_volt [λt. Cx V_s - V_{out} t; (λt. -V_{out} t)]
[&1 - semi_switch (t - t_on); semi_switch (t - t_{on}] t)
∧ (∀t. ∼(t = t_{on}) ⇒
kcl [ind_curr (λt. V_L t) L i_o; cap_curr C (λt. -V_{out} t);
res_curr R (λt. -V_{out} t)] t)

In the above definition, V_s is the supply voltage, V_{out} is the voltage drop at the junction of all these components, with respect to the ground, and V_L is the voltage drop across the inductor. However, due to the the presence of the switching junction, we model the inductor voltage, in the first conjunct, using the `switch_volt` function, which is provided with two lists; one for all the possible voltage drops, and the other with all the corresponding switching functions for every mode, and an independent variable t. Where, t_{on}, is the exact switching instant. This voltage is then used to apply the conventional Kirchoff's current law, using the function kcl, which accepts a list of currents, and an independent variable, i.e., t.

This implementation model results in the ordinary linear differential equations of the system, which can be described using Definitions 7 and 8 as:

Definition 14: $\vdash \forall$ i_o V_s V_{out} L C R t_{on} t.
buck_diff_equ i_o V_s V_{out} L C R t_{on} t =
if (t $<$ t_{on}) then diff_eq_lhs $[\frac{1}{LC};\frac{1}{RC};$ 1] (V_{out}(t)) t =
diff_eq_rhs $[\frac{V_s}{LC}]$ [1] t
else diff_eq_lhs $[\frac{1}{LC};\frac{1}{RC};$ 1] (V_{out}(t)) t = diff_eq_rhs [0] [0] t

According to the proposed methodology, as a first step, we formally verify the implementation and behavior of the Buck converter using the formal model of switching function technique and linear order differential equations as:

Theorem 4: $\vdash \forall$ i_0 V_s V_L V_{out} L C R t_{on} T_p t .
A1: ($\forall$ t. V_L continuous_on $[0,t]$ $\wedge$
A2: $\sim$ (C = 0) $\wedge$
A3: (t $\in$ (0, T_p)) $\wedge$
A4: $\sim$(t = t_{on}) $\wedge$ **A5**: (t_{on} $\in$ (0, T_p)) $\wedge$
A6: ($\forall$ t. differentiable_n_vec_deri 1 V_{out} t) $\wedge$
A7: buck_ckt_impl i_0 L C R V_s V_{out} V_L t_{on} t
$\Rightarrow$ buck_diff_equ i_0 V_s V_{out} L C R t_{on} t

Assumption A1 ensures that the converter is operating in the continuous conduction mode. Assumption A2 prevents a division by zero case in the formal analysis. Assumptions A3-A4 ensure that the time is over one time period of the system and does not include the singularities, at $t_0 = 0$, $t = t_{on}$ and $t = T_p$, due to switching action. Whereas, Assumptions A5 specifies that the switching time, $t = t_{on}$, lies within the open interval defined by the single time period of the circuit. Assumption A6 formally specifies the differentiability of the function, V_{out}, and its first derivative. The predicate differentiable_n_vec_deri accepts a number, n, and function, f, and specifies the differentiability of the function upto its n^{th}-derivative. Finally, Assumption A7 specifies the formal implementation of the power converter circuit using Definition 13. The formal proof of Theorem 4 involves taking derivative of Assumption A7, which consists of piecewise functions, by employing Theorem 1.

Following the proposed methodology, the next task is to formally verify the correctness of the solution of the ordinary linear differential equations of the Buck converter in HOL-Light. Therefore, we define the piecewise solution, i.e., Equation (10), of the Buck converter in higher-order logic as:

Definition 15: ⊢ ∀ V_s c_1 c_2 c_3 c_4 s_1 s_2 t_{on} t.
  solution V_s c_1 c_2 c_3 c_4 s_1 s_2 t_{on} t =
  linear_sol [c_1; c_2] (cexp_list [s_1; s_2]) t *
             Cx (semi_switch (t - t_{on})) +
  linear_sol [c_3; c_4] (cexp_list [s_1; s_2]) t *
            Cx (&1 - semi_switch (t - t_{on})

Where V_s is the supply voltage, c_1, c_2, c_3 and c_4 are arbitrary constants, s_1 and s_2 are the roots of homogeneous differential equations corresponding to Equations (7) and (8), respectively. Whereas, the cexp_list function is a higher-order-logic function to express the exponential form of the solution for real and distinct roots, i.e., s_1 and s_2, of the circuit. It is defined as:

Definition 16: ⊢ ∀ x. (cexp_list [] = []) ∧
  cexp_list (CONS s t) = CONS (λx. cexp (s * Cx (x))) (cexp_list t)

Next, using Definition 15, we formally verify the correctness of the solution of the differential equations, in each mode of the converter, in HOL-Light as:

Theorem 5: ⊢ ∀ i_0 V_s V_{out} L C R c_1 c_2 c_3 c_4 s_1 s_2 t_{on} T_p t .
  A1: (∀ t. ∼(t = t_{on}) ⇒ V_{out} = solution V_s c_1 c_2 c_3 c_4 s_1 s_2 t_{on} t) ∧
  A2: $(s_1 = -\frac{1}{2RC} + \frac{1}{2}\sqrt{\frac{1}{(RC)^2} - \frac{4}{LC}})$ ∧
  A3: $(s_2 = -\frac{1}{2RC} - \frac{1}{2}\sqrt{\frac{1}{(RC)^2} - \frac{4}{LC}})$ ∧
  A4: $(4\ R^2\ C \leq L)$ ∧
  A5: (0 < L) ∧
  A6: (0 < R) ∧
  A7: (0 < C) ∧
  A8: (t ∈ (0, T_p)) ∧
  A9: ∼(t = t_{on}) ∧
  A10: (t_{on} ∈ (0, T_p))
      ⇒ buck_diff_equ i_0 V_s V_{out} L C R t_{on} t

Assumption A1 formally defines the output voltage V_{out} as a piecewise function, over the time period, T_p, of the converter circuit. Assumptions A2-A3 formally specify the roots of the equation. Assumption A4 formally specifies the condition on the circuit parameters for real and distinct roots. Assumptions A5-A7, ensure the positive values of inductance, resistance and capacitance of the circuit. Assumptions A8-A9 ensure that the time is over one time period of the system and does not include the singularities, at $t_0 = 0$, $t = t_{on}$ and $t = T_p$, due to switching action. Whereas, Assumptions A10 specifies that the switching time, $t = t_{on}$, lies within the open interval defined by the single time period of the circuit.

The formal verification of Theorem 5 utilized the formally verified results of Theorems 1 and 3.

Finally, we present the formally verified results of periodic steady-state voltage of of the DC-DC Buck converter as:

Theorem 6: $\vdash \forall\ \mathtt{V_s\ V_{out}\ c_1\ c_2\ c_3\ c_4\ s_1\ s_2\ t_{on}\ t\ T_p}$.

A1: $(\mathtt{t} \in (0,\ \mathtt{T_p}))\quad \wedge$

A2: $\sim(\mathtt{t} = \mathtt{t_{on}})\quad \wedge$

A3: $(\mathtt{t_{on}} \in (0,\ \mathtt{T_p}))\ \wedge$

A4: $(\forall\ \mathtt{t}.\ \sim(\mathtt{t} = \mathtt{t_{on}}) \Rightarrow \mathtt{V_{out}} = \mathtt{solution\ V_s\ c_1\ c_2\ c_3\ c_4\ s_1\ s_2\ t_{on}\ t})\ \wedge$

A5: $(\forall\ \mathtt{t.\ \ n\_vec\_deri\ 1}\ (\lambda\ \mathtt{t.\ \ V_{out}\ t)\ continuous\ at\ t})\ \wedge$

A6: $\sim(\ \mathtt{s_2} - \mathtt{s_1} = 0)\quad \wedge$

A7: $\mathtt{steady\_state\ 1\ V_{out}\ t} \Rightarrow$

$$\Bigg(\mathtt{V_{out}}(0) = \left(\frac{\mathtt{s_2}}{\mathtt{s_2} - \mathtt{s_1}}\right)\Bigg[\ \Bigg(\ \mathtt{V_{out}}(0) + \frac{1}{\mathtt{s_2}}\frac{\mathtt{d}}{\mathtt{dt}}\mathtt{V_{out}}(0) - \mathtt{V_s}\Bigg)\ \mathtt{e}^{-\mathtt{t_{on}s_1}} + \mathtt{V_s}\Bigg]\ \mathtt{e}^{-\mathtt{T_p s_1}}\ +$$

$$\left(\frac{\mathtt{s_1}}{\mathtt{s_2} - \mathtt{s_1}}\right)\Bigg[\ \Bigg(\ \mathtt{-V_{out}}(0) - \frac{1}{\mathtt{s_1}}\frac{\mathtt{d}}{\mathtt{dt}}\mathtt{V_{out}}(0) + \mathtt{V_s}\Bigg)\ \mathtt{e}^{-\mathtt{t_{on}s_1}} - \mathtt{V_s}\Bigg]\ \mathtt{e}^{-\mathtt{T_p s_2}}\Bigg)\ \wedge$$

$$\Bigg(-\frac{\mathtt{d}}{\mathtt{dt}}\mathtt{V_{out}}(0) = \left(\frac{\mathtt{s_1 s_2}}{\mathtt{s_2} - \mathtt{s_1}}\right)\Bigg[\ \Bigg(\ \mathtt{V_{out}}(0) + \frac{1}{\mathtt{s_2}}\frac{\mathtt{d}}{\mathtt{dt}}\mathtt{V_{out}}(0) - \mathtt{V_s}\Bigg)\ \mathtt{e}^{-\mathtt{t_{on}s_1}} + \mathtt{V_s}\Bigg]$$

$$\mathtt{e}^{-\mathtt{T_p s_1}} + \left(\frac{\mathtt{s_1 s_2}}{\mathtt{s_2} - \mathtt{s_1}}\right)\Bigg[\ \Bigg(\ \mathtt{-V_{out}}(0) - \frac{1}{\mathtt{s_1}}\frac{\mathtt{d}}{\mathtt{dt}}\mathtt{V_{out}}(0) + \mathtt{V_s}\Bigg)\ \mathtt{e}^{-\mathtt{t_{on}s_1}} - \mathtt{V_s}\Bigg]\ \mathtt{e}^{-\mathtt{T_p s_2}}\Bigg)$$

Assumptions A1 and A2 formally specify the analysis over one time period with singularities, at $\mathtt{t} = 0$, $\mathtt{t} = \mathtt{t_{on}}$ and $\mathtt{t} = \mathtt{T_p}$, excluded. Whereas, Assumptions A3 specifies that the switching time, $\mathtt{t} = \mathtt{t_{on}}$, lies within the open interval defined by the single time period of the circuit. Assumption A4 formally defines the output voltage $\mathtt{V_{out}}$ as a piecewise function, over the time period, $\mathtt{T_p}$, of the converter circuit. Assumption A5 formally specifies the continuity of the function and its derivative, to ensure the continuous conduction mode. Assumption A6 prevents the division by zero case in the analysis, and finally, Assumption A7 defines the steady-state of the buck converter.

The formal proof of Theorem 6 essentially consists of finding the values of the function and its derivative at $\mathtt{t} = 0$ and $\mathtt{t} = \mathtt{T_p}$, in limit sense, and the values of arbitrary constants $\mathtt{c_1}$, $\mathtt{c_2}$, $\mathtt{c_3}$ and $\mathtt{c_4}$ by utilizing the continuity assumption A5 and the one-sided limits concepts due to singularities at $\mathtt{t} = 0$, $\mathtt{t} = \mathtt{t_{on}}$ and $\mathtt{t} = \mathtt{T_p}$, due to switching action. More details about the proof can be found at [2].

The proposed foundational formalization of switching function technique and linear differential equations allowed us to formally specify and verify the nonlinear behavior of the DC-DC Buck converters in a very straightforward manner. Theorem 4 verifies that the implementation and behavior of the Buck converter by explicitly specifying the conditions on the piecewise functions, e.g., voltages in the case of DC-DC Buck converter, in the continuous conduction operating mode of the converter.

The formally verified result is very helpful in the topology selection of the converter, which is usually the first step in the design procedure and, in practice, consists of an intuitive selection of topology for a given design specification. Moreover, Theorem 5 formally verifies the correction of the solution of the linear order differential equations representing the power converter behavior. This result plays a vital role in the performance evaluation. Once the implementation and behavior (Theorem 4), and the solution (Theorem 5) of the DC-DC Buck converter is formally verified, then Theorem 6 formally verifies the relationship among different parameters of the circuit, such as voltage and circuit components, in periodic steady-state. This result is instrumental in formal verification of the design objectives, such as desired voltage levels and component values, of the circuit. However, unlike traditional techniques these formally verified results give exact conditions in terms of the parameters of the Buck converter as they have been formally verified using a sound theorem prover. Moreover, these results are generic in terms of universally quantified variables and contain an exhaustive set of assumptions required for the validity of the results.

6 Conclusion

In this paper, we presented a formal methodology to conduct the formal time-domain based periodic steady-state analysis of power converters. The power converters are characterized by the switching functionality, which imparts to the structural changes of the converter circuit and a nonlinear mathematical analysis. To model the structural changes in the circuit, we developed the formal model of the circuit analysis technique, called switching function technique, and also developed a formal model of linear differential equations to formally specify the behavior of the converters. To cater for the nonlinearities in the analysis, the integral property of the Heaviside step function as a generalized function is verified. This logical formalism is then applied to the DC-DC Buck converter to formally verify the implementation and behavior of the converter's circuit, solution of its linear ordinary differential equations in all modes of the converter's circuit and the steady-state voltage relationship of the DC-DC Buck converter.

The proposed formalization can be extended to incorporate the formal small-signal modeling analysis of the power converters. Moreover, the formalization is based upon the complex valued functions to formally analyze the periodic steady-state analysis of power converters, which are characterized by the discontinuity due to switching action, therefore, the formalization is also equally applicable to analyze many other discontinuous phenomenon ubiquitous in many fields of Physics and engineering.

References

[1] Milton Abramowitz and Irene A. Stegun. *Handbook of mathematical functions: with formulas, graphs, and mathematical tables*, volume 55. Courier Corporation, 1964.

[2] Asad Ahmed. Formal periodic steady-state analysis of power converters in time-domain. `http://save.seecs.nust.edu.pk/projects/fpssapc/`. [Online; accessed 1-March-2019].

[3] Achraf Ben Amar, Ammar B. Kouki, and Hung Cao. Power approaches for implantable medical devices. *Sensors*, 15(11):28889–28914, 2015.

[4] B. Jayant Baliga. *Fundamentals of power semiconductor devices.* Springer Science & Business Media, 2010.

[5] Soumitro Banerjee and George C. Verghese. *Nonlinear phenomena in power electronics.* Wiley-IEEE Press, 2001.

[6] Sidi Mohamed Beillahi, Umair Siddique, and Sofiène Tahar. Formal analysis of power electronic systems. In *International Conference on Formal Engineering Methods*, pages 270–286. Springer, 2015.

[7] Philip J. Davis and Philip Rabinowitz. *Methods of numerical integration.* Courier Corporation, 2007.

[8] Manjusha Dawande, Victor Donescu, Ziwen Yao, and V. Rajagopalan. Recent advances in simulation of power electronics converter systems. *Sadhana*, 22(6):689–704, 1997.

[9] Ali Emadi. *Handbook of automotive power electronics and motor drives.* CRC Press, 2017.

[10] Robert W. Erickson and Dragan Maksimovic. *Fundamentals of power electronics.* Springer Science & Business Media, 2007.

[11] John Harrison. HOL Light: An overview. In *International Conference on Theorem Proving in Higher Order Logics*, pages 60–66. Springer, 2009.

[12] John Harrison. The HOL Light theory of Euclidean space. *Journal of Automated Reasoning*, 50(2):173–190, 2013.

[13] M. David Kankam and Malik E. Elbuluk. A survey of power electronics applications in aerospace technologies. `https://ntrs.nasa.gov/archive/nasa/casi.ntrs.nasa.gov/20020013943.pdf`, 2001. [Online; accessed 1-March-2019].

[14] Ram P. Kanwal. *Generalized functions: Theory and technique.* Springer Science & Business Media, 2012.

[15] Tuo Yeong Lee. *Henstock-Kurzweil integration on Euclidean spaces*, volume 12. World Scientific, 2011.

[16] Dragan Maksimovic. Automated steady-state analysis of switching power converters using a general-purpose simulation tool. In *Power Electronics Specialists Conference*, volume 2, pages 1352–1358. IEEE, 1997.

[17] Christos C. Marouchos. *The switching function: Analysis of power electronic circuits*, volume 17. The Institution of Engineering and Technology(IET), 2006.

[18] Marcia Verônica Costa Miranda and Antônio Marcus Nogueira Lima. Formal verifica-

tion and controller redesign of power electronic converters. In *Industrial Electronics, IEEE International Symposium on*, volume 2, pages 907–912, May 2004.

[19] Muhammad Usman Sanwal and Osman Hasan. Formally analyzing continuous aspects of cyber-physical systems modeled by homogeneous linear differential equations. In *International Workshop on Design, Modeling, and Evaluation of Cyber Physical Systems*, pages 132–146. Springer, 2015.

[20] Matthew Senesky, Gabriel Eirea, and Tak-John Koo. Hybrid modelling and control of power electronics. In *International Workshop on Hybrid Systems: Computation and Control*, pages 450–465. Springer, 2003.

[21] David R. Stoutemyer. Crimes and misdemeanors in the computer algebra trade. *Notices of the American Mathematical Society*, 38(7):778–785, 1991.

[22] Thomas G. Wilson. Life after the schematic: The impact of circuit operation on the physical realization of electronic power supplies. *Proceedings of the IEEE*, 76(4):325–334, 1988.

Received 13 August 2018

Probabilistic Analysis of Dynamic Fault Trees using HOL Theorem Proving

Yassmeen Elderhalli, Waqar Ahmad, Osman Hasan, Sofiène Tahar
Electrical and Computer Engineering
Concordia University, Montreal, QC, Canada
{y_elderh, waqar, o_hasan, tahar}@ece.concordia.ca

Abstract

Dynamic Fault Trees (DFTs) is a widely used failure modeling technique that allows capturing the dynamic failure characteristics of systems in a very effective manner. Simulation and model checking have been traditionally used for the probabilistic analysis of DFTs. Simulation is usually based on sampling and thus its results are not guaranteed to be complete, whereas model checking employs computer arithmetic and numerical algorithms to compute the exact values of probabilities, which contain many round-off errors. Leveraging upon the expressive and sound nature of higher-order-logic (HOL) theorem proving, we propose, in this paper, a formalization of DFT gates and their probabilistic behaviors as well as some of their simplification properties in HOL based on the algebraic approach. This formalization would allow us to conduct the probabilistic analysis of DFTs by verifying generic mathematical expressions about their behavior in HOL. In particular, we formalize the AND, OR, Priority-AND, Functional DEPendency, Hot SPare, Cold SPare and the Warm SPare gates and also verify their corresponding probabilistic expressions in HOL. Moreover, we formally verify an important property, $Pr(X < Y)$, using the Lebesgue integral as this relationship allows us to reason about the probabilistic properties of the Priority-AND gate and the *Before* operator in HOL theorem proving. We also formalize the notion of conditional densities in order to formally verify the probabilistic expressions of the Cold SPare and the Warm SPare gates. In order to illustrate the usefulness of our formalization, we use it to formally analyze the DFT of a Cardiac Assist System.

1 Introduction

A Fault Tree (FT) [22] represents an effective way of graphically modeling the causes of failure in a system in the form of a rooted failure tree. A typical FT consists of a *top event* representing system failure, basic failure events modeling the components failure and the FT gates, which combine the basic failure events and allow components failure to propagate to the top event. FTs are categorized as: Static FTs (SFTs) and Dynamic FTs (DFTs). SFTs capture the causes of failure in a system without considering the failure dependencies or sequences between the system components. DFTs, on the other hand, capture the failure dependencies in systems, which represent a more realistic approach to model the behavior of real-world systems.

Fault Tree Analysis (FTA) can be used to examine the failure characteristics of the given system qualitatively and quantitatively. In the former analysis, the combinations and sequences of basic failure events, associated with the system components, are determined in the form of cut sets and cut sequences [22]. While the quantitative analysis allows estimating the failure probability of the system based on component's failure probabilities among other metrics. Usually, Markov chain (MC) based analysis or algebraic approaches are used to perform DFT analysis. In the Markov chain based analysis, the DFT is first converted into its equivalent MC and then the analysis is conducted on the resulting MC. Complex systems often lead to a MC with a large number of states. The MCs of such complex systems can be analyzed using a modularization approach that divides the corresponding FT into SFT and DFT parts [19]. The SFT part is analyzed using traditional combinatorial analysis methods, such as Binary Decision Diagrams (BDDs) [22], while the DFT part is analyzed using MCs [23]. This kind of modularization approach has been implemented in the Galileo tool [24]. In the algebraic approach, an algebra similar to the ordinary Boolean algebra is used to reduce the structure function (expression) of the *top event* of the DFT [12]. This reduced expression is then used to derive the failure probability of the given system based on the failure probabilities of DFT gates.

Traditionally, DFTs are either analyzed by analytically deriving the system failure probability expression or using computer-based simulation tools. In the former method, firstly cut-sequences consisting of basic failure events are obtained and then the probabilistic Principle of Inclusion-Exclusion (PIE) [12] is used to manually derive the probability of failure of the overall system. This kind of manual manipulation is prone to human errors and can produce erroneous results especially when dealing with large DFTs. The latter method is more extensively used due to its scalability and user friendliness. Several simulation tools are available that provide GUI editors that obtain the system FT model from the user and return the analysis results based on the assigned failure distribution to the system components at a given instant of time. However, simulation cannot be guaranteed to produce complete and accurate results due to the involvement of numerical techniques, such as Monte Carlo simulation [17], and pseudo random variables. Due to the above-mentioned inaccuracies, both analytical and simulation based methods are not suitable to conduct the failure analysis of safety-critical systems.

As an accurate alternative, formal methods have been recently utilized for analyzing FTs. Probabilistic model checkers (PMC), such as STORM [6], have been used to perform the quantitative analysis of DFTs [9]. However, due to the state-based nature of PMCs, they cannot be used to verify generic expressions for probability of failure. In addition, their usage is only limited to exponential distributions, which in the context of reliability analysis, for example, do not consider the aging of systems components. Due to the sound nature of higher-order-logic (HOL) theorem proving, it has been successfully used to formalize basic SFT gates [1], which have been in turn used to conduct the SFT-based analysis of several systems, including an air traffic management system [2]. However, this formalization is only limited to SFTs. So far, there is no formalization in HOL that supports the probabilistic failure analysis of DFTs. Recently, we have presented a hybrid methodology based on both interactive theorem proving and model checking for formal analysis of DFTs [8]. The main idea is to first conduct the qualitative analysis of a given DFT, based on the algebraic approach [12], using theorem proving and then quantitatively analyze the simplified DFT model using the STORM model checker. Since a PMC is involved in estimating the probabilities quantitatively, this methodology cannot provide generic expressions for probability and its usage is only limited to exponential distributions. Moreover, the formal definitions of DFT gates in [8] cannot cater for conducting the probabilistic analysis using HOL theorem proving as the behavior of the DFT gates has been captured using numbers instead of random variables.

In order to perform the complete probabilistic analysis of DFTs within a higher-order-logic theorem prover by verifying generic expressions of probability of failure, we propose to improve our formalization of the DFT gates in higher-order logic that uses the algebraic approach, presented in [12], as its foundation. The choice to formalize an algebraic approach to conduct the DFT analysis is motivated by the fact that HOL is well known for modeling systems that can be mathematically expressed. In addition, using HOL theorem proving we can formally verify generic expressions that cannot be obtained and verified using other formal tools and the algebraic approach fits perfectly with these features of HOL theorem proving. The foremost task in this work is to identify an algebraic approach to formalize DFTs so that the formal DFT analysis can be constructed within a theorem prover. In this respect, we have to consider the availability of foundational theories, like measure and probability, and their compatibility with the chosen approach. We identified the algebraic approach, initially proposed by Merle [12], to formalize DFTs among other options (e.g., [18]). Despite the fact that the presented formalization is based on an existing algebraic approach [12], it bears its own research challenges. For example, it is well-known that a theorem proving based proof requires many intricate proof guidance and explicit reasoning that a mathematician doing a paper based proof would sometimes ignore. Thus, in our experience with the formalization of the algebraic approach, we also had to take many modeling decisions, choose appropriate data types, identify missing assumptions for the validity of results, devise proof strategies and verify many helper theorems to facilitate the process of formal DFT analysis in a theorem prover. The paper highlights these details while we present our formalization. Based on this novel formalization, we also formally verify the DFT algebraic reduction properties. Then, using the available probability theory formalization [13], we also formally verify the failure probability relationships of all commonly used DFT gates, i.e., AND, OR, Priority-AND (PAND), Functional DEPendency (FDEP), Hot SPare gate (HSP), Cold SPare gate (CSP) and Warm SPare gate (WSP). In order to verify the failure probability relationship of some of these DFT gates, we are required to formalize the $Pr(X < Y)$ describing the effect of one system component failing before the other or one after the other. This property is mainly verified by using Lebesgue integral properties [15, 21]. In addition, we formalize the notion of conditional density functions, which is necessary to formally verify the probabilistic relationships of the spare gates. The HOL4 theorem prover [10] was a natural choice for this formalization as it has the required theories such as: the probability theory and the Lebesgue integral [15]. In addition, we use the existing formalization of the probabilistic PIE in HOL4 [1]. The abovementioned formalizations can be utilized to conduct the DFT-based failure analysis of a variety of real-world systems within the sound core of a theorem prover. For illustration purposes, we present the formal DFT-based failure analysis of a Cardiac Assist System (CAS) [5], which is a safety-critical DFT benchmark. We first reduce the original structure function of the system's top event using the formally verified simplification theorems. Then, we utilize the probabilistic PIE [1] to formally verify a generic failure probability expression of the Cardiac Assist System whereas the failure characteristics of its components are represented as generic probability distribution and density functions.

1.1 Contributions of the Paper

The main contributions of the paper are summarized as:

- Providing a framework for the probabilistic analysis of DFT within a theorem prover, which offers a sound and rigorous method for conducting DFT analysis by providing formally verified generic expressions of probability of failure.

- Development of reasoning steps for the verification of DFT gate properties, which, to the best of our knowledge, are not available in the literature or even in [12].
- Providing the formal definitions of DFT gates, which are somewhat different than the expressions provided in [12].
- Verifying a generic expression for the probabilistic failure behavior of a cardiac assist system in HOL theorem proving, which involves identifying the required conditions for the generic expression to hold.

1.2 Paper Organization

The rest of the paper is structured as follows: Section 2 presents some preliminaries about the probability theory and the Lebesgue integral in HOL4 that will facilitate the understanding of the rest of the paper. In Section 3, we present our HOL formalization of DFT gates and the corresponding simplification properties. Section 4 provides the verification details of the probabilistic behavior of the DFT gates. Section 5 presents the formalization of the probabilistic failure behavior of the Cardiac Assist System. Finally, we conclude the paper in Section 6.

2 Preliminaries

In this section, we present some preliminaries that are required for the understanding of the proposed formalization.

2.1 Probability Theory

The probability theory is formalized based on the measure theory in HOL4 [15]. A measurable space is represented as a pair $(\mathcal{X}, \mathcal{A})$, where $\mathcal{X}$ represents a space and $\mathcal{A}$ a set of measurable sets. The functions `space` and `subsets` are defined in HOL to return $\mathcal{X}$ and $\mathcal{A}$, respectively of a measurable space $(\mathcal{X}, \mathcal{A})$. A measure is generally a function that designates a certain number to a set, which represents the size of this set [13]. It is defined as the triplet $(\mathcal{X}, \mathcal{A}, \mu)$, where $\mathcal{X}$ represents the space, $\mathcal{A}$ represents the measurable sets and finally μ represents the measure. Three functions, `m_space, measurable_sets` and `measure`, are defined in HOL to return the space $(\mathcal{X})$, measurable sets $(\mathcal{A})$ and measure (μ) of a measure space, respectively [16]. A probability space is defined as a measure space, with the added condition that the probability measure for the entire space is equal to 1.

Random variables are formalized as measurable functions that map events from the probability space to some other σ- algebra space s. Random variables are defined in HOL4 as [13]:

Definition 2.1.

```
⊢ ∀X p s.  random_variable X p s ⇔
           prob_space p ∧ X ∈ measurable (p_space p, events p) s
```

where `prob_space p` ensures that `p` is a probability space with `p_space` as its space and `events` as its measurable sets. `X ∈ measurable (p_space p, events p) s` ensures that `X` belongs to the set of measurable functions from the probability space `p` to σ-algebra space `s` [16]. Measurable spaces `s` and `(p_space p, events p)` are ensured to be σ-algebra spaces using the `measurable` function.

The probability distribution of a random variable X represents the probability that the random variable X belongs to a set A. This is equivalent to finding the probability of the event $\{X \in A\}$, which can also be represented using the preimage as $X^{-1}(A)$. The probability distribution is defined in HOL4 as [13]:

Definition 2.2.

```
⊢ ∀p X. distribution p X = (λs.  prob p (PREIMAGE X s ∩ p_space p))
```

where `s` is a set of elements of the space that the random variable X maps to. For a random variable that maps the probability space (p) into another measurable space, the push forward measure is a measure that uses the space and subsets of the measurable space as its space and measurable sets and uses the distribution of the random variable as its measure part [11]. In general, the push forward measure for any measurable function X from measure M to measure N can be expressed as:

Definition 2.3.

```
⊢ ∀ M N f.  distr M N f =
    (m_space N, measurable_sets N,
        λA. measure M (PREIMAGE f A ∩ m_space M))
```

A density measure is used to define a density function, f, over the measure space M as [11]:

Definition 2.4.

```
⊢ ∀ M f.  density M f =
    (m_space M, measurable_sets M,
         λA. pos_fn_integral M (λ x.  f x * indicator_fn A x ) )
```

where `pos_fn_integral` represents the Lebesgue integral of positive functions as will be described in the following section.

The cumulative distribution function (CDF) of a random variable X is usually used when we are interested in finding the probability that the random variable is less than or equal to a certain value. It is formally defined for real values as [1]:

Definition 2.5.

```
⊢ ∀p X t.  CDF p X t = distribution p X {y | y ≤ (t:real)}
```

It is worth mentioning that the CDF can be defined for extended-real (`extreal`) random variables as well, where `extreal` is a HOL data-type that includes the real numbers beside ±∞. However, in our formalization we will use the CDF of real random variables, as it is required to integrate their density functions over the real line.

When dealing with multiple random variables, the probabilistic *Principle of Inclusion and Exclusion* (PIE) provides a very interesting relationship between the probability of the union of different events. It can be expressed as:

$$Pr(\bigcup_{i=1}^{n} A_i) = \sum_{t\neq\{\},t\subseteq\{1,2,\ldots,m\}} (-1)^{|t|+1} Pr(\bigcap_{j\in t} A_j) \qquad (1)$$

It has been formally verified in HOL4 as follows [1]:

Theorem 2.1.

```
⊢ ∀p L.
    prob_space p ∧ (∀ x.  MEM x L ⇒ x ∈ events p) ⇒
    (prob p (union_list L = sum_set {t | t ⊆ set L ∧ t ≠ {}}
    (λt.  -1 pow (CARD t+1) * prob p (BIGINTER t))
```

where L is the list of events that we are interested in expressing the probability of their union.

In order to be able to handle multiple random variables, a pair measure (often called binary product measure) is required to be able to model joint distribution measures. This pair measure can be used also in a nested way to model the joint distribution measure of multiple random variables. The pair measure is defined as the product of two measures. It was initially formalized in Isabelle/HOL [11] and was then ported to HOL4 [20]. The space and the measurable sets of this pair measure are generated using the Cartesian product of the spaces and the measurable sets of the participating measures, while the measure part is defined using the Lebesgue integral.

Since there are real and extended-real data-types in HOL4, there exist two Borel spaces, one over the real line (`borel`) [21] and the second over the extended-real line (`Borel`) [14]. The Lebesgue-Borel measure is required to integrate over the real line. In particular, we need the Lebesgue-Borel measure in this work to integrate the density functions of the random variables over the real line. The Lebesgue-Borel measure is a measure defined over the real line, which uses the real line as its space and the Borel sets as its measurable sets. The Lebesgue-Borel measure is defined in HOL4 as `lborel`, which uses the real borel sigma algebra (`borel`) generated by the open sets of the real line as well as the Lebesgue measure [21].

The independence of random variables is an important property when dealing with multiple random variables. In general, for any two random variables X and Y, the probability of the intersection of their events is equal to the multiplication of the probability of the individual events. The independence of random variables is defined as `indep_vars` [20]:

Definition 2.6.

```
⊢ indep_vars p M X ii =
   (∀i.  i ∈ ii ⇒
      random_variable (X i) p
        (m_space (M i), measurable_sets (M i))) ∧
   indep_sets p
      (λi.  {PREIMAGE f  A ∩ p_space p |
         (f = X i) ∧ A ∈ measurable_sets (M i)}) ii
```

where p is the probability space and M is the measure space that the random variable X maps to. In this case, M and X are indexed by a number from the set of numbers ii, which gives the possibility of defining the independence for multiple random variables that map from the probability space to different spaces. The function indep_vars defines the independence by first ensuring that the group of input functions X are random variables and that their event sets are independent using indep_sets. Using indep_sets, the probability of the intersection of any sub-group of events of the random variables is equal to the multiplication of the probability of the individual events.

Using indep_vars, the independence of two random variables is defined as [20]:

Definition 2.7.

```
⊢ indep_var p M_x X M_y Y =
  indep_vars p (λi.  if i = 0 then M_x else M_y)
               (λi.  if i = 0 then X else Y) {x | (x = 0) ∨ (x = 1)}
```

We define several functions that facilitate handling our formalization. The first function is measurable_CDF, which is defined as:

Definition 2.8.

```
⊢ ∀p X. measurable_CDF p X = (λx.  CDF p X x) ∈ measurable borel Borel
```

This function ensures that the CDF of random variable X is measurable from the borel space to the Borel space. In other words, it ensures that the CDF is measurable from the real line to the extended-real line. This implies that the domain for this CDF is the real line and the range is the extended-real line.

We define another function, cont_CDF, which ensures that the CDF is continuous. It is formally defined as:

Definition 2.9.

```
⊢ ∀p X. cont_CDF p X = ∀z.  (λx.  real (CDF p X x)) contl z
```

where the function real typecasts the value of CDF from extended-real to real data-type, and contl ascertains that the function is continuous over all values in its domain. It is worth mentioning that X is a real valued random variable. However, the CDF returns extended-real. As the continuity of functions is defined in HOL4 for real valued functions, it is required to typecast the value of the CDF from extended-real to real. In addition, since the values of the CDF range from 0 to 1, as it represents a probability, this function is the same in both cases but with different datatypes. Therefore, if the function is continuous in the extended-real, then it is continuous using the real datatype. Furthermore, later we will use extended-real random variables, therefore, it is required to typecast their values using the `real` function.

Next, we define a function, `rv_gt0_ninfinity`, to ensure that the input random variables of a DFT can only have the range $[0, +\infty)$:

Definition 2.10.

```
⊢ (rv_gt0_ninfinity [] = T) ∧
  (rv_gt0_ninfinity (h::t) = (∀s.  0 ≤ h s ∧ h s ≠ PosInf) ∧
     (rv_gt0_ninfinity t))
```

Finally, we define a function, `den_gt0_ninfinity` to ensure the proper values for the marginal, joint and conditional density functions:

Definition 2.11.

```
⊢ ∀f_xy f_y f_cond.
    den_gt0_ninfinity f_xy f_y f_cond ⇔
    ∀x y.
      0 ≤ f_xy (x,y) ∧ 0 < f_y y ∧ f_y y ≠ PosInf ∧ 0 ≤ f_cond y x
```

where `f_xy` is the joint density function, `f_y` is the marginal density function, and finally `f_cond` is the conditional density function of X given Y. This function can be used to assign the mentioned conditions to other functions and not necessarily only the density functions.

2.2 Lebesgue Integral

The Lebesgue integral is defined in HOL4 using positive simple functions, which are measurable functions defined as a linear combinations of indicator functions of measurable sets representing a partition of the space X [15]. A positive simple function, g, can be represented using the triplet $(s,\ a,\ x)$ as [15]:

$$\forall t \in X,\ g(t) = \sum_{i \in s} x_i \mathbf{1}_{a_i}(t), \qquad x_i \geq 0 \tag{2}$$

where s is a finite set of partition tags, x_i is a sequence of positive `extreal` numbers, a_i is a sequence of measurable sets and $\mathbf{1}_{a_i}$ is the indicator function of measurable set a_i and is defined as [15]:

Definition 2.12.

```
⊢ ∀A. indicator_fn A = (λx.  if x ∈ A then 1 else 0)
```

The Lebesgue integral is first defined for positive simple functions and then extended for positive functions for measure μ as [14]:

$$\int_X f d\mu = sup\{\int_X g\ d\mu \mid g\ \leq\ f\ and\ g\ positive\ simple\ function\} \tag{3}$$

It is usually required that the probability of an event for a random variable to be expressed using the integration of the random variable's distribution. This is verified in HOL4 as [13]:

Theorem 2.2.

```
⊢ ∀X p s A.
    random_variable X p s ∧ A ∈ subsets s ⇒
    (distribution p X A =
     integral (space s, subsets s, distribution p X)(indicator_fn A))
```

In the above theorem, X can be a continuous or discrete random variable. However, in our DFT formalization, we are only interested in continuous random variables as they represent the time of failure of system components.

3 Formalization of Dynamic Fault Trees in HOL

Our previous formalization of DFT gates and operators was based on the algebraic approach [12], where the DFT events are treated based on their time of occurrence (failure of corresponding components) [8]. However, these formal definitions cannot cater for the probabilistic analysis of system failures, which is the scope of the current paper. Therefore, we provide an improved formalization of DFT gates and operators using functions of time that can be represented as random variables when carrying out the formal probabilistic analysis of the given DFT based on the algebraic approach presented in [12]. However, there are some missing gaps in the paper-and-pencil proofs available in [12] that we were able to fill using our formalization, particularly that we had to build our formalization on top of some existing HOL theories, such as the Lebesgue integral and probability theories. In [12], there is no direct description on how to build the DFT analysis based on the above-mentioned theories. Besides this, we also had to use different strategies for some proofs. All these differences will be highlighted throughout Sections 3 and 4.

3.1 Identity Elements and Temporal Operators

Similar to ordinary Boolean algebra, the DFT algebraic approach defines identity elements that are important in the simplification process of the DFT [12]. The DFT identity elements are: the *ALWAYS* element representing an event that always occurs (fails) from time 0 and the *NEVER* element, which describes an event that never occurs (fails). The formal definitions of these elements are shown in Table 1, where `PosInf` represents $+\infty$ in HOL4. We define the time of failure of the events as lambda abstracted functions that accept an arbitrary data-type that represents an element from the probability space and return the time. so that they can be later treated as random variables. For example, the time of failure of a component is a random variable X and can be expressed in lambda abstraction form as `(λs. X s)`.

Temporal operators are also required to model the DFT gates in the algebraic approach [12]. These operators are: *Before* ($\lhd$), *Simultaneous* (Δ) and *Inclusive Before* ($\unlhd$). Each one of these operators accepts two inputs, which can be subtrees or basic events that represent faults of system components. The output event of the operator occurs according to a certain sequence of occurrence for the input events, i.e., the time of occurrence of the first (left) input is less than, equal to or less than or equal to the occurrence time of the second input (right) for the *Before*, the *Simultaneous* and the *Inclusive Before* operators, respectively. The time of occurrence of the output event of all operators is equal to the time of occurrence of the first input event (left). The mathematical expressions of these operators as well as their corresponding HOL formalizations are shown in Table 1, where `X` and `Y` represent the time of occurrence of events X and Y, respectively.

It is worth mentioning that if the inputs of the *Simultaneous* operator are basic events with continuous failure distributions, then the output of this operator can never fail [12]. This is because the time of failure is continuous, and the possibility that two system components failing at the same time can be neglected. As a consequence, it is assumed in the algebraic approach that any two *different* basic events

Table 1: Definitions of Identity Elements and Temporal Operators

Element/Operator	Mathematical Expression	Formalization
`Always element`	$d(ALWAYS) = 0$	`⊢ ALWAYS = (λs. (0:extreal))`
`Never element`	$d(NEVER) = +\infty$	`⊢ NEVER = (λs. PosInf)`

`Before`	$d(X\lhd Y)=\begin{cases} d(X), & d(X) < d(Y) \\ +\infty, & d(X) \geq d(Y) \end{cases}$	`⊢ ∀X Y. D_BEFORE X Y =`
		`(λs. if X s < Y s then X s else PosInf)`

`Simultaneous`	$d(X\Delta Y)=\begin{cases} d(X), & d(X) = d(Y) \\ +\infty, & d(X) \neq d(Y) \end{cases}$	`⊢ ∀X Y. D_SIMULT X Y =`
		`(λs. if X s = Y s then X s else PosInf)`

`Inclusive Before`	$d(X\unlhd Y)=\begin{cases} d(X), & d(X) \leq d(Y) \\ +\infty, & d(X) > d(Y) \end{cases}$	`⊢ ∀ X Y. D_INCLUSIVE_BEFORE X Y =`
		`(λs. if X s ≤ Y s then X s else PosInf)`

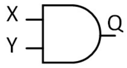
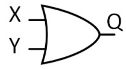
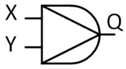
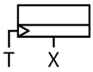
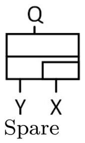
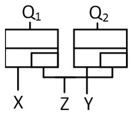
can never fail at the same time. This can be expressed for basic failure events of the inputs of the given DFT as [12]:

$$d(X \Delta Y) = NEVER \tag{4}$$

3.2 Formalization of FT Gates and Simplification Theorems

Our formalization of all FT gates; static and dynamic, and their mathematical expressions [12] are presented in Table 2.

Table 2: DFT Gates

Gate	Mathematical Expression	Formalization
 AND	$d(X \cdot Y) = max(d(X), d(Y))$	⊢ ∀X Y. D_AND X Y = (λs. max (X s)(Y s))
 OR	$d(X + Y) = min(d(X), d(Y))$	⊢ ∀X Y. D_OR X Y = (λs. min (X s)(Y s))
 PAND	$d(Q_{PAND}) = \begin{cases} d(Y), & d(X) \leq d(Y) \\ +\infty, & d(X) > d(Y) \end{cases}$	⊢ ∀X Y. PAND X Y = (λs. if X s ≤ Y s then Y s else PosInf)
 FDEP	$d(X_T) = min(d(X), d(T))$	⊢ ∀X T. FDEP X T = (λs. min (X s)(T s))
 Spare	$d(Q_{CSP}) = \begin{cases} d(X), & d(Y) < d(X) \\ +\infty, & d(Y) \geq d(X) \end{cases}$	⊢ ∀X Y. CSP Y X = (λs. if Y s < X s then X s else PosInf)
	$d(Q_{HSP}) = max(d(Y), d(X))$	⊢ ∀X Y. HSP Y X = (λs. max (Y s)(X s))
	$d(Q_{WSP}) = d(Y \cdot (X_d \lhd Y) + X_a \cdot (Y \lhd X_a) + Y \Delta X_a + Y \Delta X_d$	⊢ ∀Y X_a X_d. WSP Y X_a X_d = D_OR(D_OR(D_OR (D_AND Y (D_BEFORE X_d Y)) (D_AND X_a (D_BEFORE Y X_a))) (D_SIMULT Y X_a))(D_SIMULT Y X_d)
 Shared Spare	$d(Q_1) = d(X \cdot (Z_d \lhd X) + Z_a \cdot (X \lhd Z_a) + X \cdot (Y \lhd X))$	⊢ ∀X Y Z_a Z_d. shared_spare X Y Z_a Z_d = D_OR (D_OR (D_AND X (D_BEFORE Z_d X)) (D_AND Z_a (D_BEFORE X Z_a))) (D_AND X (D_BEFORE Y X)))

3.2.1 AND and OR Gates

The AND ($\cdot$) and OR ($+$) gates can be modeled based on the time of occurrence of their output events. For the AND gate, the output occurs when both of its input events occur and the time of occurrence of the output is modeled as the maximum time of occurrence of both input events [12]. For the OR gate, the output occurs once one of its input events occurs. Therefore, we formalize it as the minimum time of occurrence of the inputs [12]. In Table 2, `max` and `min` are the HOL4 functions that represent the maximum and the minimum functions, respectively. It is important to notice that we define the AND and OR gates as lambda abstracted functions that accept two inputs that are also functions. This would enable defining the inputs later as random variables to represent the time of failure function of system components. This also applies to the formal definitions of the rest of DFT gates.

3.2.2 Priority AND Gate (PAND)

The PAND gate, shown in Table 2, captures the sequence of occurrence (failure) of its inputs. The output event of this gate occurs if all input events occur in a certain sequence (conventionally from left to right). In Table 2, we provide both the mathematical and formal definitions of the PAND gate. Then we verify that the behavior of the PAND can also be represented using the temporal operators as [12]:

$$Q = Y \cdot (X \trianglelefteq Y) \tag{5}$$

We verify the above relationship in HOL4 as follows:

Theorem 3.1. ⊢ `∀X Y. PAND X Y = D_AND Y (D_INCLUSIVE_BEFORE X Y)`

This result ascertains that the behavior of PAND gate is correctly captured in our formal definition. It is worth mentioning that in [12] the PAND gate is defined as Equation (5). However, we define it using a mathematical expression as in Table 2, which represents its actual behavior, and then verify that this definition is equal to the definition provided in [12] as in Theorem 3.1.

3.2.3 Functional DEPdency Gate (FDEP)

The FDEP is used to model the dependencies in the failure behavior between the system components. In other words, it is used when the failure of one component triggers the failure of another. For the FDEP gate, shown in Table 2, event X can occur if it is triggered by the failure of T or if it occurs by itself. As a result, the occurrence time of X_T (triggered X) equals the minimum time of occurrence of T and X. From the FDEP definition, we can notice that its behavior is equivalent to the behavior of the OR gate.

3.2.4 Spare Gates

Modeling spare parts in real systems is necessary when analyzing the probability of failure of the overall system, as these spares are used to replace the main parts after their failure. The main part Y of the spare gate, shown in Table 2, is replaced by the spare part X after a failure occurs. The spare gate has three variants depending on the type of the spare:

- ***Cold SPare Gate (CSP)***: The spare part can only fail while it is active.
- ***Hot SPare Gate (HSP)***: The spare part can fail in both the active and the dormant states with the same probability.
- ***Warm SPare Gate (WSP)***: The spare part can fail in both the dormant and active states with different probabilities.

While manipulating the structure function of the DFT, it is required to distinguish between the two states of the spare part, i.e., the active state and the dormant state, therefore a different variable is assigned to each state. For example, for the spare gate in Table 2, variable X is assigned X_d and X_a for the dormant and active states, respectively [12]. This is required in case of a WSP gate, where the spare part has two different states. Recall that in the case of a CSP gate, it is not necessary to use these subscripts, since the spare part in the CSP gate does not work in the dormant state. Therefore, the active state only affects the DFT behavior and is included in the expressions. In the HSP gate, the spare part has the same behavior for both states and no subscript is required to distinguish between these two.

It can be noticed from the definition of the WSP gate that the output of the spare occurs in two cases; if the spare fails in its dormant state, then the main part fails or the main part fails then the spare is activated and then it fails in its active state. The last two terms in the WSP definition cover the possibility that the spare and the main part fail at the same time. This can happen if the main part and the spare are functionally dependent on the same trigger. The WSP represents the general case for the spare gates, while the CSP and HSP represent special cases of the WSP, where the spare cannot fail or is fully functioning in its dormant state. We have defined mathematical expressions for both the CSP gate for basic events and the HSP gate to facilitate using their expressions in DFT analysis. However, as will be seen shortly, we have verified that the behavior of our expressions is equivalent to a WSP under certain conditions. For the CSP gate, the output occurs if the main part fails then the spare is activated and then the spare fails while it is active. Since the spare part of the HSP has the same failure distribution in both of its states, the output of the HSP occurs when both inputs (main and spare) fail. Therefore, its behavior is equivalent to an AND gate. We formally verify that the WSP gate is equivalent to an HSP gate when the spare part in its dormant state is equal to its active state.

Theorem 3.2. ⊢ `∀X Y. WSP Y X X = HSP Y X`

Moreover, we formally verify that the WSP gate is equivalent to a CSP gate, if the spare part cannot fail in its dormant state. We formally verify this as:

Theorem 3.3. ⊢ `∀X_a X_d Y. (X_d = NEVER) ∧`
`(∀s.  ALL_DISTINCT [Y s; X_a s]) ⇒ WSP Y X_a X_d = CSP Y X_a`

where `X_d = NEVER` indicates that the spare part cannot fail in its dormant state, and `ALL_DISTINCT` ensures that the inputs cannot fail at the same time. This is because we defined the CSP gate for basic events. As can be seen from the above theorem, the CSP gate only deals with the active state of the spare, therefore, when dealing with a CSP there is no need to use the subscript.

In some real-world applications, a spare part can replace one of two main parts. This case is represented using shared spare gates as shown in Table 2 [8]. The expression of the output Q_1 of the first gate is listed in Table 2 [12]. This expression implies that the output Q_1 of this gate occurs in three different situations: *(i)* if the main part X fails, then the spare fails while it is active (Z_a), *(ii)* if the spare part fails in its dormant state Z_d, then the main part fails, or *(iii)* if the second main part (of the other gate) Y fails before X, and thus the spare is not available to replace X when it fails. We use the DFT operators to model the behavior of this gate, as shown in Table 2.

In the DFT algebraic approach, many simplification theorems exist and are used to reduce the structure function of the top event [12]. In [8], we verified over 80 simplification theorems. However, these theorems were based on our old definitions of the DFT gates and operators that cannot cater for probabilistic analysis. We verify all these theorems for the new definitions, presented in this paper, and the details can be accessed from [7]. These simplification theorems range from simple ones, such as commutativity of the AND, OR and Simultaneous operator, to more complex ones that include combinations of all the operators. Table 3 includes some of these verified properties.

Table 3: Examples of Formally Verified Simplification Theorems

DFT Algebra Theorems	HOL Theorems
$X{+}Y{=}Y{+}X$	⊢ ∀X Y. D_OR X Y = D_OR Y X
$X.NEVER{=}NEVER$	⊢ ∀X. D_AND X NEVER = NEVER
$X{\lhd}(Y{+}Z){=}(X{\lhd}Y).(X{\lhd}Z)$	⊢ ∀ X Y Z. D_BEFORE X (D_OR Y Z) = D_AND (D_BEFORE X Y)(D_BEFORE X Z)
$X{\trianglelefteq}(Y{+}Z){=}(X{\trianglelefteq}Y).(X{\trianglelefteq}Z)$	⊢ ∀ X Y Z. D_INCLUSIVE_BEFORE X (D_OR Y Z) = D_AND (D_INCLUSIVE_BEFORE X Y) (D_INCLUSIVE_BEFORE X Z)
$(X{\trianglelefteq}Y){+}(X\Delta Y){=}X{\trianglelefteq}Y$	⊢ ∀X Y. D_OR (D_INCLUSIVE_BEFORE X Y) (D_SIMULT X Y) = D_INCLUSIVE_BEFORE X Y

4 Formal Verification of DFT Probabilistic Behavior

In order to formally verify the probability of failure of the top event of a DFT, it is required to formally model and verify the probability of failure expression for each DFT gate. We assume that the basic events of the DFT are independent. However, in some cases these events can be dependent; in particular in the case of CSP and WSP, where the failure of the main part affects the operation and failure of the spare part. We handle this by first introducing the probabilistic behavior of the gates for independent events, then we present the probabilistic behavior of the WSP and the CSP gates, which deal with dependent events. At the end of this section, we present a summary of the challenges that we faced during the formalization of the probabilistic failure behavior of DFT gates.

4.1 Probabilistic Behavior of Gates with Independent Events

Assuming that we are interested in finding the probability of failure until time t, the following four expressions can be used to express the probability of any DFT gate with independent basic events [12]:

$$Pr\{X \cdot Y\}(t) = F_X(t) \times F_Y(t) \tag{6a}$$

$$Pr\{X + Y\}(t) = F_X(t) + F_Y(t) - F_X(t) \times F_Y(t) \tag{6b}$$

$$Pr\{Y \cdot (X \lhd Y)\}(t) = \int_0^t f_Y(y)\ F_X(y)\ dy \tag{6c}$$

$$Pr\{X \lhd Y\}(t) = \int_0^t f_X(x)(1 - F_Y(x))\ dx \tag{6d}$$

where F_X and F_Y represent the CDFs of the random variables X and Y, respectively, and f_X and f_Y represent their corresponding PDFs.

Equation (6a) represents the probability of the AND and HSP gates, which results from the probability of intersection of two independent events. Equation (6b) describes the probability of the OR and FDEP gates, which corresponds to the probability of union of two independent events. Equation (6c) represents the probability of having two basic events occurring in sequence one *after* the other until time t, i.e., $Pr(X < Y)$ until time t or $Pr(X < Y \wedge Y \leq t)$, which is the failure probability of the PAND for basic events. Finally, the probability of the *Before* operator is represented by Equation (6d), which is the probability of having event X occurring *before* event Y until time t, i.e., $Pr(X < Y \wedge X \leq t)$. The difference between the last two events (*before* and *after*) is that in the *before* event, we are just interested in finding the probability of failure of X until time t with the condition that X fails before Y. So, it is not necessary that Y fails. While in the *after* event, we find the probability of failure of Y until time t with the condition that Y fails after X. So, it is required that both X and Y fail in sequence.

Since the probability is applied for sets that belong to the events of the probability space, we define a `DFT_event` that satisfies the condition that the input function is less than or equal to time t, which represents the moment of time until which we are interested in finding the probability of failure. Without this `DFT_event`, there is no possible way to apply the probability directly to DFT gates. We first need to create the `DFT_event` for the time-to-failure function of the output event of any gate or DFT, then apply the probability to it.

Definition 4.1.

```
⊢ ∀p X t.  DFT_event p X t = {s | X s ≤ Normal t} ∩ p_space p
```

where `Normal` typecasts the type of t from `real` to `extreal`, p represents the probability space and X represents the time-to-failure function.

We formally verify the equivalence between the probability of the `DFT_event` of an extended real function and its equivalent CDF of the real version of the function as:

Theorem 4.1.

```
⊢ ∀X p t.  (∀s.  X s ≠ PosInf ∧ 0 ≤ X s) ⇒
    (CDF p (λs.  real (X s)) t = prob p (DFT_event p X t))
```

where `real` is mirror opposite to the typecasting `Normal` operator. This typecasting is required as the `DFT_event` is defined for `extreal` data-type, and the CDF is defined for real random variables only. Therefore, it is required to ensure that the input function does not equal $+\infty$ and is greater than or equal to 0 since it represents the time of failure of a system component.

4.1.1 Probabilistic Behavior of AND, HSP, OR and FDEP Gates

To formally verify Equations (6a) and (6b), we verify the equivalence of the DFT event of the AND gate to the intersection of two events and the OR as the union:

Lemma 4.1.

```
⊢ ∀p t X Y.
    DFT_event p (D_AND X Y) t = DFT_event p X t ∩ DFT_event p Y t
```

Lemma 4.2.

```
⊢ ∀p t X Y.
    DFT_event p (D_OR X Y) t = DFT_event p X t ∪ DFT_event p Y t
```

Based on the independence of random variables and using Theorem 4.1, we formally verify Equation (6a) in HOL4 as:

Theorem 4.2.

```
⊢ ∀p t X Y. rv_gt0_ninfinity [X; Y] ∧
    indep_var p lborel (λs.  real (X s)) lborel (λs.  real (Y s)) ⇒
    (prob p (DFT_event p (D_AND X Y) t) =
     CDF p (λs.  real (X s)) t * CDF p (λs.  real (Y s)) t
```

where `indep_var` ensures the independence of the random variables, X and Y, over the Lebesgue-Borel (`lborel`) measure [20]. `rv_gt0_ninfinity` is required since we are dealing with the real versions of the random variables. It is a logical condition, since any real-world component will eventually fail, so we are interested only in dealing with the time of failure that is not ∞.

In Theorem 4.2, the random variables are type-casted as real-valued, using the operator `real`, to function over the Lebesgue-Borel (`lborel`) measure. `lborel` is purposely used here to facilitate the Lebesgue integration over the real line when expressing the probabilities of the *before* and *after* events. Theorem 4.2 represents the probability of the AND gate and the HSP gate, since the behavior of the HSP is equivalent to the behavior of the AND gate.

We formally verify Equation (6b) based on the probabilistic PIE and the independence of random variables and using Theorem 4.1 as:

Theorem 4.3.

```
⊢ ∀p t X Y. rv_gt0_ninfinity [X; Y] ∧
    All_distinct_events p [X;Y] t ∧
    indep_var p lborel (λs.  real (X s)) lborel (λs.  real (Y s)) ⇒
    (prob p (DFT_event p (D_OR X Y) t) =
     CDF p (λs.  real (X s)) t + CDF p (λs.  real (X s)) t -
     CDF p (λs.  real (X s)) t × CDF p (λs.  real (Y s)) t)
```

where `All_distinct_events` ascertains that the event sets are not equal. We formally define it as:

Definition 4.2.

```
⊢ All_distinct_events p L t =
  ALL_DISTINCT (MAP (λx.  DFT_event p x t) L
```

where `ALL_DISTINCT` is a HOL4 predicate, which ensures that the elements of its input list are not equal, `MAP` is a function that applies the input function `(λx. DFT_event p x t)` to all the elements in the list `L` and returns a list. This condition is required for the probabilistic PIE.

Theorem 4.3 provides the probability of the OR gate as well as the FDEP gate, since the behavior of the FDEP is equivalent to the OR gate.

It is worth noting that in [12], Equations (6a) and (6b) were just presented without any information on how to link them to the definitions of the AND and OR gates. We should recall that the AND and OR gates are defined as the maximum and minimum of their operands. Looking at these definitions does not give any knowledge about how the probability of the AND gate is equivalent to the probability of the intersection or how the probability of the OR gate is equal to the probability of the union. However, using our formalization and utilizing our formal definition of `DFT_event`, we are able to verify that the `DFT_event` of the AND gate is equal to the intersection of the input events and that the `DFT_event` of the OR gate is equal to the union of the input events. Based on this, we can ensure that the probability of the AND and OR gates are represented using Equations (6a) and (6b), respectively.

4.1.2 Probabilistic Behavior of PAND Gate and Before Operator

We verify Equations (6c) and (6d) as Theorems 4.4 and 4.5, respectively.

Theorem 4.4.

```
⊢ ∀X Y p fy t.
    rv_gt0_ninfinity [X; Y] ∧ 0 ≤ t ∧ prob_space p ∧
    indep_var p lborel (λs.  real (X s)) lborel (λs.  real (Y s)) ∧
    distributed p lborel (λs.  real (Y s)) fy ∧ (∀y.  0 ≤ fy y) ∧
    cont_CDF p (λs.  real (X s)) ∧
    measurable_CDF p (λs.  real (X s)) ⇒
    (prob p (DFT_event p (Y·(X◁Y)) t) =
     pos_fn_integral lborel
        (λy.  fy y *
             (indicator_fn {w | 0 ≤ w ∧ w ≤ t} y *
              CDF p (λs.  real (X s)) y)))
```

Theorem 4.5.

```
⊢ ∀X Y p fy t.
    rv_gt0_ninfinity [X; Y] ∧ 0 ≤ t ∧ prob_space p ∧
    indep_var p lborel (λs.  real (X s)) lborel (λs.  real (Y s)) ∧
    distributed p lborel (λs.  real (X s)) fx ∧ (∀x.  0 ≤ fx x) ∧
    measurable_CDF p (λ s.  real (Y s)) ⇒
    (prob p (DFT_event p (X ◁ Y) t) =
     pos_fn_integral lborel
        (λx.  fx x *
             (indicator_fn {u | 0 ≤ u ∧ u ≤ t} x *
              (1- CDF p (λs real (Y s)) x)))
```

where `pos_fn_integral` is the Lebesgue integral for positive functions [15], `fy` and `fx` are the PDF of random variables of the real version of functions Y and X, respectively. `cont_CDF` is required in Theorem 4.4 as we need to prove that $Pr(X \leq t)$ and

$Pr(X < t)$ are equal, and this is not valid unless the CDF function is continuous (`cont`).

Verifying Theorems 4.4 and 4.5 is not a straightforward task due to the involvement of Lebesgue integration. To the best of our knowledge, this is the first time that these proofs are formally verified in a theorem prover, where we are able to identify the exact steps to reach the final form of Theorems 4.4 and 4.5. In addition, in [12], Equation (6c) is presented without any proof, while a proof is presented for Equation (6d) that is based mainly on the probability of disjoint events and utilizes derivatives to reach the final expression. However, we have been able to verify the same expression of Equation (6d), but following a different and simpler proof, which is similar to the proof of Equation (6c) to reach the final form of Theorem 4.5 without using derivatives. We first prove the probability of sets of real random variables in the form of integration before extending the proofs to extended real functions.

Proof Strategy for Theorem 4.4

To verify Theorem 4.4, we first express the event set that corresponds to the integration in Equation (6c) as:

$$(X, Y)^{-1}\{(u, w) \mid u \ < \ w \ \wedge \ 0 \leq \ w \wedge \ w \ \leq \ t\} \tag{7}$$

Then we verify that the probability of this set can be written using the integration as in Equation (6c). Therefore, we verify the relationship between the distribution and the integration of positive functions using the push forward measure (`distr`):

Theorem 4.6.

```
⊢ ∀X p M A.
    measure_space M ∧
    random_variable X p (m_space M, measurable_sets M) ∧
    A ∈ measurable_sets M ⇒
    (distribution p X A =
     pos_fn_integral (distr p M X) (indicator_fn A))
```

It is worth mentioning that this theorem can be used in the verification process of other applications and not only for DFT analysis. We use Theorem 4.6 to verify the relationship between the probability and the integration of the joint distribution F_{XY} of two independent random variables as:

$$Pr(X, Y)^{-1}(A) = \int \mathbf{1}_A \ dF_{XY} \tag{8}$$

We formalize this relationship in HOL4 and use a property, which converts the distribution of a pair measure of independent measures into the pair measure of the individual distributions [20], to split the integral of joint distributions into two integrals of the individual distributions ($\int\int \mathbf{1}_A dF_X dF_Y$). In order to reach the final form of Equation (6c), we express it in the form of two integrals:

$$\int_0^t f_Y(y) \times F_X(y)\ dy = \int_0^t \int_{-\infty}^{y} f_Y(y) \times f_X(x)\ dx\ dy \tag{9a}$$

$$= \int_0^t f_Y(y) \Big(\int_{-\infty}^{y} f_X(x)\ dx\ \Big) dy \tag{9b}$$

The problem in Equations (9a) and (9b) lies in the fact that the outer integral is a function of the inner integral, i.e., for the inner integral we are integrating until y which is the variable of the outer integral. To be able to handle this formally, we verify that the indicator function of the set in Equation (7) can be written in the form of the multiplication of two indicator functions, where one is a function of the other.

Lemma 4.3.

```
⊢ ∀x y t.
    indicator_fn {(u,w) | u < w ∧ 0 ≤ w ∧ w ≤ t}(x,y) =
    indicator_fn {w| 0 ≤ w ∧ w ≤ t} y * indicator_fn {u|u < y} x
```

In order to use the above-mentioned lemma and the set on the left hand side, we need to verify that this set is measurable in the two dimensional borel space, i.e., the set belongs to the measurable sets of `pair_measure lborel lborel`. This property can be verified based on the fact that the countable union of measurable sets is also measurable. We verify this fact on the rational numbers $\mathbb{Q}$ as follows:

Theorem 4.7.

```
⊢ ∀m s.
    measure_space m ∧ (∀n.  n ∈ Q_set ⇒ s n ∈ measurable_sets m) ⇒
    BIGUNION (IMAGE s Q_set) ∈ measurable_sets m
```

where m in our case is `pair_measure lborel lborel`. This theorem is generic and can be used in other contexts than DFTs.

The purpose of using the set of rational numbers is that we need a countable set that can be used to express the set in Lemma 4.3 as the union of borel rectangles. We verify this in HOL4 as:

Lemma 4.4.

```
⊢ ∀t.  BIGUNION
          {{u | u < real q} × {w | real q < w ∧ 0 ≤ w ∧ w ≤ t} |
            q ∈ Q_set} =
       {(u,w) | u < w ∧ 0 ≤ w ∧ w ≤ t}
```

Besides this, we also verify a lemma that the sets in the union of Lemma 4.4 are measurable sets in the `pair_measure lborel lborel` as:

Lemma 4.5.

```
⊢ ∀t q.  {u | u < real q} × {w | real q < w ∧ 0 ≤ w ∧ w ≤ t} ∈
         measurable_sets (pair_measure lborel lborel)
```

We can use the proof steps of the previous lemmas to verify the same properties for similar sets, which is essential for other gates expressions. This facilitates dealing with other events in the future, by following the steps in our proof.

By using the above lemmas, we can reason that the original set is a measurable set in the `pair_measure lborel lborel` as:

Lemma 4.6.

```
⊢ ∀t.  {(u,w) | u < w ∧ 0 ≤ w ∧ w ≤ t} ∈
       measurable_sets (pair_measure lborel lborel)
```

We use Lemmas 4.3 and 4.6 to verify that the expression given in Equation (9b) is equal to $\int_A dF_X dF_Y$, where A is the set that specifies the boundaries of the integral. We verify this in HOL4 using the push forward measure as:

Lemma 4.7.

```
⊢ ∀X Y p t.
    prob_space p ∧ indep_var p lborel X lborel Y ⇒
    (pos_fn_integral (pair_measure (distr p lborel X)
             (distr p lborel Y))
       (λ(x,y).  indicator_fn{(u,w) |u < w ∧ 0 ≤ w ∧ w ≤ t }(x,y) =
     pos_fn_integral (distr p lborel Y)
        (λy.  indicator_fn {w|0 ≤ w ∧ w ≤ t} y *
          pos_fn_integral(distr p lborel X)
             (λx.  indicator_fn {u | u < y} x)))
```

where `pair_measure (distr p lborel X) (distr p lborel Y)` represents the joint distribution of the push forward measures of random variables X and Y over the borel space.

We verify several essential properties for CDF in order to prove that the inner integral of Lemma 4.7 is equal to $F_X(y)$ or formally to `(CDF p X y)`. In order to have the PDF of random variable Y in the integral, we assume that the random variable Y has a PDF by defining a density measure for Y. We ported the following definition, `distributed`, from Isabelle/HOL [11], where f in this definition is the PDF of random variable X, and the measure part of the density measure is the integral of this PDF. Using this definition, the integral of f is equal to the distribution of the random variable X.

Definition 4.3.

```
⊢ ∀p M X f.
    distributed p M X f ⇔
    X ∈
    measurable(m_space p,measurable_sets p)
      (m_space M,measurable_sets M) ∧
    f ∈ measurable(m_space M,measurable_sets M) Borel ∧
    AE M {x | 0 ≤ f x} ∧ (distr p M X = density M f)
```

where `density` is the density measure, and `AE M {x | 0 ≤ f x }` ensures that the PDF `f` is almost everywhere positive over the measure `M`. We also use a theorem that replaces the integration with respect to the density measure by the PDF with respect to the original measure (`lborel` in our case) [11]. In addition to the previously verified theorems, we also prove some additional properties, such as sigma finite measure for the push forward measure over the borel space (`sigma_finite_measure (distr p lborel X)`). We also verify that the space generated by the pair measure of two distributions over the borel space is sigma algebra (`sigma_algebra (m_space (pair_measure (distr p lborel X)(distr p lborel Y)), measurable_sets (pair_measure (distr p lborel X)(distr p lborel Y)))`). Moreover, we verify that the space generated by the space and the measurable sets of the pair measure of `lborel` is also a sigma algebra (`sigma_algebra (m_space (pair_measure lborel lborel), measurable_sets (pair_measure lborel lborel))`). Finally, we prove that the set of the left-hand side of Equation (6c) is equal to the set that corresponds to the integration of the right-hand side of the same equation as:

Lemma 4.8.

```
⊢ ∀p t X Y.
    rv_gt0_ninfinity [X; Y] ∧ 0 ≤ t ⇒
    (DFT_event p (Y·(X◁Y)) t =
     PREIMAGE (λx.  (real (X x), real (Y x)))
        {(u,w) | u < w ∧ 0 ≤ w ∧ w ≤ t} ∩ p_space p
```

Based on all the above mentioned lemmas, we are able to verify the original goal for Equation (6c) as in Theorem 4.4.

Proof Strategy for Theorem 4.5

For the verification of Theorem 4.5, we follow almost the same steps for the previous proof. We start by first writing the event set for the integration as:

$$(X,Y)^{-1}\{(u,w) \mid 0 \ \leq \ u \ \wedge \ u \leq \ t \ \wedge \ u \ < \ w \ \} \tag{10}$$

Then, we describe the indicator function of this set as the multiplication of two indicator functions as:

Lemma 4.9.

```
⊢ ∀x y t.
    indicator_fn {(u,w) | 0 ≤ u ∧ u ≤ t ∧ u < w}(x,y) =
    indicator_fn {u | 0 ≤ u ∧ u ≤ t} x * indicator_fn {w | x < w} y
```

Similar to the procedure, explained previously for the set of the after event in Lemmas 4.4, 4.5 and 4.6, we verify that the set of the before event is a measurable set in the `pair_measure lborel lborel`.

Finally, we rewrite Equation (6d) as:

$$\begin{aligned} Pr\{X \lhd Y\}(t) &= \int_0^t \int_x^\infty f_X(x)\ f_Y(y)\ dy\ dx \\ &= \int_0^t f_X(x)\ \Big(\int_x^\infty f_Y(y)\ dy\ \Big) dx \end{aligned} \tag{11}$$

We verify some additional properties for the CDF in order to complete the proof. For example, we verify that $\int_x^\infty f_Y(y)\ dy$ is equal to $1 - F_Y(x)$. Similarly, we also formally verify that the event of the left-hand side of Equation (6d) is equal to the set that corresponds to the integration of the right-hand side of the same equation. We use the set in Equation (10) to verify this as:

Lemma 4.10.

```
⊢ ∀p t X Y.
    rv_gt0_ninfinity [X; Y] ∧ 0 ≤ t ⇒
    (DFT_event p (X◁Y) t =
     PREIMAGE (λs.  (real (X s),real (Y s)))
        {(u,w) | 0 ≤ u ∧ u < w ∧ u ≤ t} ∩ p_space p
```

Based on all these verified theorems, we are able to formally verify Theorem 4.5.

So far, we presented the formal verification of the probabilistic behavior of:

- The AND and HSP gates using Theorem 4.2 (since they are equivalent).
- The probability of the OR and FDEP gates using Theorem 4.3 (since they are equivalent).
- The probability of the PAND gate for basic events using Theorem 4.4.
- The probability of the *Before* operator using Theorem 4.5.

There is no probability of failure for the *Simultaneous* operator as it is eliminated for basic events according to Equation (4). This implies that the probability of the *Inclusive Before* operator is equal to the probability of the *Before* operator for basic events.

4.2 Probabilistic Behavior of Gates with Dependent Events

The probabilistic behavior of the CSP and WSP requires dealing with dependent events, as the failure of the main part affects the behavior of the spare part. Therefore, it is required to approach the proof in a different manner.

For the CSP, the failure distribution of the spare part is affected by the failure time of the main part, as the cold spare starts working after the failure of the main part. Hence, the failure distribution of the spare part is dependent on the failure of the main part. The probability of failure for the output event of a CSP with Y as the main part and X as the spare part is given by [12]:

$$Pr(Q_{CSP})(t) = \int_0^t \Big(\int_v^t f_{(X_a|Y=v)}(u)du \Big) f_Y(v)dv \tag{12}$$

where $f_{(X_a|Y=v)}$ is the conditional probability density function for the spare part in its active state (X_a) given that the main part(Y) has failed at time v. As mentioned previously, the subscript of X_a can be omitted, since the spare part of the CSP gate does not work in its dormant state and we are only concerned with the active state, so using X directly with CSP means that we are dealing with the active state and not the dormant one. It can be noticed from Equation (12) that the failure distribution of the spare part is affected by the failure of the main part. Hence, these two input events are not independent, and we cannot utilize the previously verified relationships in Section 4.1 to verify the probabilistic behavior of the CSP gate.

For the WSP gate with two basic events, the output fails in two cases, Case 1: when the main part fails, then the spare fails in its active state (this case is similar to the CSP case); Case 2: when the spare part fails in its dormant state, then the main part fails with no spare to replace it. In the latter case, the failure distribution of the spare part in its dormant state is independent of the main part. Hence, we can use the previously verified expressions for this case. The probability expression for a WSP with X as the spare part (X_a for the active state and X_d for the dormant state) and Y as the main part is expressed as [12]:

$$Pr(Q_{WSP})(t) = \int_0^t \Big(\int_v^t f_{(X_a|Y=v)}(u)du \Big) f_Y(v)dv + \int_0^t f_Y(u)F_{X_d}(u)du \qquad (13)$$

where F_{X_d} is the CDF of X in its dormant state. The first part of Equation (13) represents the probability of a CSP and the second part represents the probability when the spare fails before the main part. For the second part, Y and X_d are considered to be independent as the failure of one of them does not affect the failure of the second and hence we can use Equation (6c) for this case.

We verify Equations (12) and (13) as Theorems 4.8 and 4.9, respectively.

Theorem 4.8.

```
⊢ ∀p X Y f_xy f_y f_cond t.
    rv_gt0_ninfinity [X; Y] ∧ 0 ≤ t ∧
    (∀y.
        cond_density lborel lborel p
          (λs.  real (X s)) (λs.  real (Y s)) y f_xy f_y f_cond) ∧
    prob_space p ∧ den_gt0_ninfinity f_xy f_y f_cond ⇒
    (prob p (DFT_event p (CSP Y X) t) =
     pos_fn_integral lborel
        (λy.
           indicator_fn {u | 0 ≤ u ∧ u ≤ t} y * f_y y *
           pos_fn_integral lborel
              (λx.  indicator_fn {w | y < w ∧ w ≤ t} x * f_cond y x )))
```

Theorem 4.9.

```
⊢ ∀p Y X_a X_d t f_y f_xy f_cond.
    prob_space p ∧ (∀s.  ALL_DISTINCT [X_a s; X_d s; Y s]) ∧
    (D_AND X_a X_d = NEVER) ∧ rv_gt0_ninfinity [X_a; X_d; Y] ∧ 0 ≤ t ∧
    (∀y.
       cond_density lborel lborel p
          (λs.  real (X_a s))(λs.  real (Y s)) y f_xy f_y f_cond) ∧
    den_gt0_ninfinity f_xy f_y f_cond ∧
    indep_var p lborel (λs.  real (X_d s)) lborel (λs.  real (Y s)) ∧
    cont_CDF p (λs.  real (X_d s)) ∧
    measurable_CDF p (λs.  real (X_d s)) ⇒
    (prob p (DFT_event p (WSP Y X_a X_d) t) =
     pos_fn_integral lborel
        (λy.
           indicator_fn {u | 0 ≤ u ∧ u ≤ t} y * f_y y *
           pos_fn_integral lborel
              (λx.  indicator_fn {w | y < w ∧ w ≤ t} x * f_cond y x ))+
     pos_fn_integral lborel
        (λy.
           f_y y *
           (indicator_fn {u | 0 ≤ u ∧ u ≤ t} y *
            CDF p (λs.  real (X_d s)) y )))
```

where p is the probability space, `f_xy` is the joint density function for X and Y, `f_y` is the marginal density function for Y, `cond_density` defines the conditional density function (`f_cond`) for X given that $(Y = y)$ and `den_gt0_ninfinity` ensures the proper values for the density functions as mentioned in Section 2.

It is noticed that the spare part in the CSP is used without any subscript, i.e., it is used as X, since the spare has only one state in the CSP, which is the active state. Therefore, there is no need to use any subscript to distinguish between the dormant and the active states. While in the WSP, we need to distinguish between the two states, i.e., active and dormant, hence the usage of X_a and X_d. For Theorem 4.9, the condition `D_AND X_a X_d = NEVER` ensures that the spare part can only fail in one of its states but not both. This condition is different from `D_SIMULT X_a X_d = NEVER`, as the former means that if one of the inputs occurs, then the other cannot occur at all. While the latter means that both inputs cannot occur at the same time, they can occur at different times. This second condition is ensured in our case using `ALL_DISTINCT`. In addition, it is assumed that the spare part in the dormant (X_d) state is independent of the main part Y since the failure of the spare part in its dormant state is not affected by the failure of the main part. As with the previous theorems in Section 4.1, we need to use the typecast operator `real` with the random variables, since the random variables are of type `extreal` and the integral over the `lborel` requires real random variables.

In [12], a proof has been introduced for the above expressions, which is based mainly on the total expectation theorem [4]. However, we have been able to conduct the same proof in a simpler manner based on conditional density functions as explained below.

Proof Strategy for Theorem 4.8 (CSP Gate)

In order to verify Theorem 4.8, we formalize the conditional density function as [3]:

Definition 4.4.

```
⊢ ∀M1 M2 p X Y y f_xy f_y f_cond.
    cond_density M1 M2 p X Y y f_xy f_y f_cond ⇔
    random_variable X p (m_space M1, measurable_sets M1) ∧
    random_variable Y p (m_space M2, measurable_sets M2) ∧
    distributed p (pair_measure M1 M2) (λx.  (X x, Y x)) f_xy ∧
    distributed p M2 Y f_y ∧ (f_cond y = (λx.  f(x,y) / f_y y))
```

where `p` is the probability space, `M1` and `M2` are the measure spaces that the random variables X and Y map to, respectively (we will use `lborel` in our case), `f_xy` is the joint density function for X and Y, `f_y` is the marginal density function of Y and finally, `f_cond` is the conditional density function of X given $(Y = y)$.

The conditional density function definition ensures that X and Y are random variables with joint density function `f_xy` and a marginal density function `f_y`. It is noticed from the definition of the conditional density function `f_cond` that it is a function of x only, and it can have different variants depending on the value of Y that we are conditioning at, i.e., y. This is why `f_cond` takes y as a parameter.

From Definition 4.4, we formally verify the following relationship between the conditional density, the joint density and the marginal density functions, given that $f_Y(y) \neq 0$:

$$f_{XY}(x,y) = f_{X|Y=y}(x) \times f_Y(y) \tag{14}$$

The above equation can be formalised in HOL4 as:

Theorem 4.10.

```
⊢ ∀M1 M2 p X Y f_cond x y f_xy f_y.
    (∀y.  f_y y ≠ 0 ∧ f_y y ≠ PosInf ∧ f_y y ≠ NegInf) ∧
    cond_density M1 M2 p X Y y f_xy f_y f_cond ⇒
    (f_xy (x,y) = f_cond y x * f_y y)
```

The condition `f_y y ≠ 0` is required, as this function will be used in the denominator of the conditional density and it cannot equal to 0. In addition, since we are dealing with extended-real numbers, `f_y y` cannot equal infinity. This theorem is applicable to any conditional density function that satisfies the given conditions.

The second step in verifying the expression of the CSP is by verifying that the probability of the joint random variables is equal to the iterated integrals of the joint density function. This can be expressed as:

$$Pr(X,Y)^{-1}(A) = \int\int \mathbf{1}_A \times f_{XY}(x,y)dx\ dy \tag{15}$$

We use Theorem 4.6 to verify this in HOL4 as:

Theorem 4.11.

```
⊢ ∀p X Y f_xy A.
    distributed p (pair_measure lborel lborel) (λx.  (X x, Y x)) f_xy ∧
    prob_space p ∧ (∀x.  0 ≤ f_xy x) ∧
    A ∈ measurable_sets (pair_measure lborel lborel)⇒
    (prob p (PREIMAGE (λx.  (X x, Y x)) A ∩ p_space p) =
     pos_fn_integral lborel
        (λy.
          pos_fn_integral lborel
             (λx.  indicator_fn A (x,y) * f_xy (x,y))))
```

Then, we express the probability of the joint random variables using the conditional density function as:

$$Pr(X,Y)^{-1}(A) = \int\int \mathbf{1}_A \times f_{(X|Y=y)}(x) \times f_Y(y)\ dx\ dy \tag{16}$$

We verify this in HOL4, using Theorems 4.10 and 4.11, as:

Theorem 4.12.

```
⊢ ∀p X Y f_xy f_y f_cond A.
    (∀y.  cond_density lborel lborel p X Y y f_xy f_y f_cond) ∧
    prob_space p ∧ (∀x.  0 ≤ f_xy x) ∧
    (∀y.  0 < f_y y ∧ f_y y ≠ PosInf) ∧
    A ∈ measurable_sets (pair_measure lborel lborel)⇒
    (prob p (PREIMAGE (λx.  (X x, Y x)) A ∩ p_space p) =
     pos_fn_integral lborel
        (λy.
          pos_fn_integral lborel
             (λx.  indicator_fn A (x,y) * f_cond y x * f_ y y )))
```

In order to be able to reach the final form of Equation (12), we need first to express the event set that corresponds to the integration in Equation (12) as:

$$(X,Y)^{-1}\{(x,y) \mid y \ < \ x \wedge \ x \ \leq \ t \ \wedge \ 0 \ \leq \ y \ \wedge \ y \ \leq \ t\} \tag{17}$$

We verify in HOL4 that this set corresponds to the `DFT_event` of the CSP gate as:

Lemma 4.11.

```
⊢ ∀X Y p t.
    rv_gt0_ninfinity [X; Y] ∧ 0 ≤ t ⇒
    (DFT_event p (CSP Y X) t =
     PREIMAGE (λs.  (real (X s), real (Y s)))
        {(x,y)| y < x ∧ x ≤ t ∧ 0 ≤ y ∧ y ≤ t} ∩ p_space p)
```

In addition, we verify that the event set in Lemma 4.11 is measurable in `pair_measure lborel lborel`. Finally, we verify that the indicator function of the set in Lemma 4.11 can be expressed as the multiplication of two indicator functions to determine the boundaries of the iterated integrals in Equation (12) as:

Lemma 4.12.

```
⊢ ∀x y t.
    indicator_fn {(w,u) | u < w ∧ w ≤ t ∧ 0 ≤ u ∧ u ≤ t} (x,y) =
    indicator_fn {w | y < w ∧ w ≤ t} x *
    indicator_fn {u | 0 ≤ u ∧ u ≤ t} y
```

Using all these verified theorems and lemmas, we formally verify Theorem 4.8.

Proof Strategy for Theorem 4.9 (WSP Gate)

For the verification of Theorem 4.9, it is evident that the probability expression involves the probability of the CSP gate in addition to the probability of the *after* expression of Theorem 4.4. Therefore, we choose to verify that the event of the WSP for basic events is equivalent to the union of two sets as:

Lemma 4.13.

```
⊢ ∀p Y X_a x_d t.
    (∀s.  0 ≤ Y s) ∧ (∀s.  ALL_DISTINCT [X_a s; X_d s; Y s]) ∧
    (D_AND X_a X_d = NEVER) ⇒
    (DFT_event p (WSP Y X_a X_d) t =
      {s | Y s < X_a s ∧ X_a s ≤ Normal t ∧
          0 ≤ Y s ∧ Y s ≤ t} ∩ p_space p ∪
      {s | X_d s < Y s ∧ Y s ≤ Normal t } ∩ p_space p)
```

Then, we verify that the above two sets are disjoint. As these two sets are disjoints then the probability of the original set is equivalent to the sum of the probabilities of the disjoint sets. Based on this, we verify that the probability of the first set (`{s | Y s < X_a s ∧ X_a s ≤ Normal t ∧ 0 ≤ Y s ∧ Y s ≤ t} ∩ p_space p`) is equal to the probability of the CSP gate, which will result in the first term in the addition of the conclusion of Theorem 4.9. We also verify that the probability of the second set in Lemma 4.13 (`{s | X_d s < Y s ∧ Y s ≤ Normal t} ∩ p_space p)`) is expressed using Theorem 4.4, which will result in the second term of the addition of the conclusion of Theorem 4.9. As a result, we have the probability of the WSP as in Theorem 4.9.

In this section, we formally verified the probabilistic behavior of the DFT gates: AND, OR, HSP, FDEP, PAND, CSP, WSP and the Before operator besides the formalization of expressions for $Pr(X < Y \wedge Y \leq t)$ and $Pr(X < Y \wedge X \leq t)$.

These verified properties are generic, i.e., universally quantified for all distribution and density functions, and can be used to formally verify the probability of failure expression of any DFT. The HOL4 proof script for this verification as well as the gate definitions is available at [7].

4.3 Summary of Formalization Challenges

In this section, we summarize the main challenges that we faced during our formalization of the DFT gates, which allows us to formally analyze DFTs in a theorem prover.

The first challenge is resolving the data-types issue. The problem in the data-types is that the gates and operators are defined as functions that return `extreal`. This is mainly required because we need to model $+\infty$ that represents the NEVER condition. However, this data-type cannot be used to represent random variables over the `lborel` measure. Any random variable defined from a probability space to the `lborel` measure should return `real` data-type. This is required because we need to integrate the density and distribution functions over the real line. Therefore, we need random variables that return `extreal` to model the gates but at the same return `real` to be used with `lborel`. We resolved this issue by using `extreal` to model the gates, but when we are conducting the probabilistic analysis we use the real version of the random variable `(λ. real (X s))`.

Secondly, after modeling the DFT and expressing the structure function of the top event using the DFT gates and operators, it is required to conduct the probabilistic failure analysis of the top event. However, the structure function cannot be used directly since it is a time-to-failure function not a set. Furthermore, in [12], there is no clear information on how to create the DFT event and link it to the structure function of the DFT top event or any other event in the fault tree. Using our formalization, we have been able to clearly and formally define a `DFT_event` that is used to create the set of moments of time until the time of failure t, as explained in Definition 4.1.

Thirdly, the probabilities of the AND and OR gates are directly presented in [12] as the probability of the intersection and union (Equations (6a) and (6b), respectively). However, the AND and the OR are defined using the maximum and minimum of their input operands, respectively. There is no information in [12] on how the AND and OR gates are related to the intersection and union of the input events. Using our formalization, we have been able to verify the relationship between the AND and the interaction of the input events utilizing our defined `DFT_event`. In a similar way, we verified the relationship between the OR gate and the union of the input events.

Another contribution is represented by introducing a formal proof in a theorem prover for the probability of failure of the PAND and Before operator, which are represented by $Pr(X < Y)$ in both forms, i.e., $Pr(X < Y \wedge Y \leq t)$ and $Pr(X < Y \wedge X \leq t)$. As mentioned earlier, the first proof of these ($Pr(X < Y \wedge Y \leq t)$) is not provided in [12], while the second one ($Pr(X < Y \wedge X \leq t)$) is presented in a different manner that involves derivatives. In our formalization, we presented, for the first time, the formal proof for $Pr(X < Y)$ in both its formats, i.e., $Pr(X < Y \wedge Y \leq t)$ that represents the probability of the PAND gate for basic events; and $Pr(X < Y \wedge X \leq t)$ that represents the probability of the before operator. In addition, we presented a formal proof for the probability of the WSP and CSP

gates based on conditional density functions, which we defined, while the proof of these gates is presented in [12] based on the law of total expectation.

Finally, while performing all of these formalizations and proofs in HOL, we identified several missing assumptions or conditions that were required to ensure the correctness of the theorems. For example, ensuring the proper values for the input random variables that represent the time-to-failure functions of the system components. These important assumptions were either unavailable in [12] or are not explicitly presented as a requirement in the final form of the theorems in [12].

It is important to highlight that the main benefit of having the formalization of DFT in higher-order logic is that it enables conducting the formal DFT analysis within the sound environment of a theorem prover, which is very useful in the context of safety-critical systems.

5 Formal Verification of the Cardiac Assist System

In order to illustrate the utilization of our formalized probabilistic behavior of the gates and operators in the last section, we present a DFT-based formal failure analysis of the Cardiac Assist System, shown in Figure 1 [5].

We first provide generic steps that can be followed in order to use our formalization of the DFTs to conduct the formal analysis of DFTs in the form of generic expressions of failure probabilities. These steps are:

1. Determine the structure function of the top event of the DFT.
2. Simplify the structure function and formally verify that the simplified version is equal to the original function obtained in step (1).
3. Create the `DFT_event` of the structure function.
4. Express the `DFT_event` of the top event as the union of multiple input events.
5. Apply the probabilistic PIE to the union of events generated in the previous step, then simplify the result of the PIE. This will result in having the summation of the probabilities of the intersection of the different events that contribute to the failure of the top event of the DFT.
6. Replace each term in the result of the PIE by its probabilistic expression based on the verified expressions in Section 4 for each gate and operator.

Step (5) requires proving many lemmas that are necessary for manipulating the result of the PIE. For example, we need to verify the associativity property of addition for a large group of numbers (in case of the Cardiac Assist system, we verified

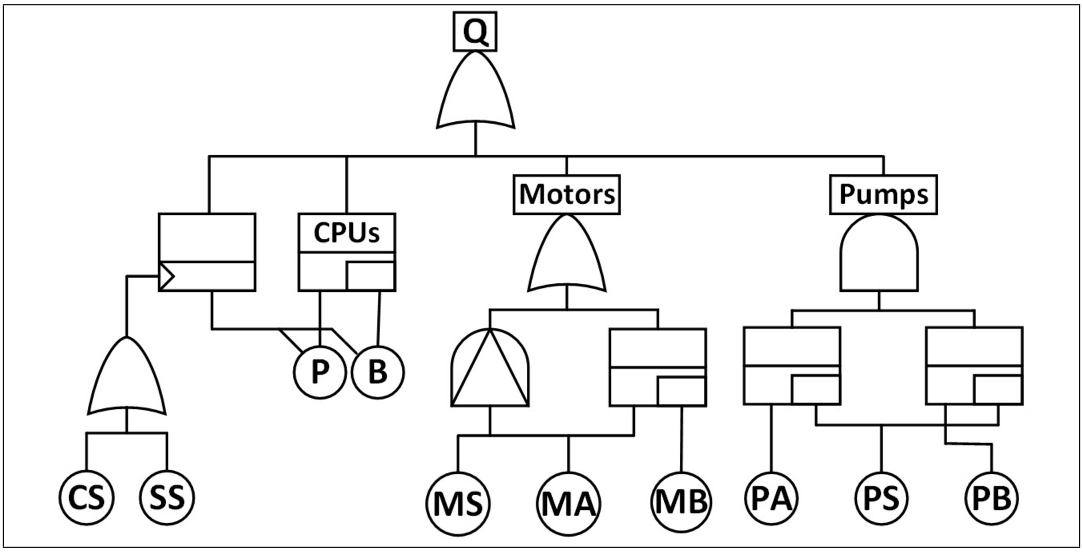


Figure 1: Cardiac Assist System

this property for 63 numbers). Although this seems a trivial task, it requires dealing with `extreal` numbers, which includes proving that for all combinations of the inputs, the result of the addition cannot equal to ∞. Step (5) also requires verifying the power set of events in a recursive way to generate a set of all combinations of the events, which is required by the PIE. Moreover, based on the independence of the input random variables, we need to verify the independence of several combinations of random variables (in the Cardiac Assist system, we verified that any two random variables out of the ten are independent, then three out of ten,... etc). We have verified these generic lemmas and they can be easily reused with other similar case studies and thus can reduce the proof efforts significantly.

In the rest of this section, we illustrate the utilization of the previous steps to perform the formal DFT analysis of the Cardiac Assist System to provide a generic expression for the probability of failure of the top event. The Cardiac Assist system consists of three main subsystems: pumps, motors and CPUs. The system has two main pumps (PA and PB) with a shared spare PS. It has three motors MS, MA, and MB, where MB replaces MA after failure. Finally, the system has one main CPU (P) and a spare CPU (B). Both CPUs are functionally dependent on a trigger, which is the union of the crossbar switch (CS) and the system supervisor (SS). In this case study, we are assuming that the spare gates are HSPs.

Our goal is to verify the probability of failure of the Cardiac Assist system by applying the probabilistic PIE considering that the input events are independent.

This can be represented mathematically as:

$$\begin{aligned} Pr(Q) =& F_{CS}(t) \ + \ F_{SS}(t) \ + \ \int_0^t f_{MA}(y) \times F_{MS}(y) \ dy \ + \\ & F_{MA}(t) \times F_{MB}(t) \ + \ F_P(t) \times F_B(t) \ + \ F_{PA}(t) \times F_{PB}(t) \times F_{PS}(t) \\ & - ... + ... - \ F_{CS}(t) \ \times F_{SS}(t) \times \Big(\int_0^t f_{MA}(y) \times F_{MS}(y) \ dy \ \Big) \\ & \times F_{MA}(t) \times F_{MB}(t) \times F_P(t) \times F_B(t) \times F_{PA}(t) \times F_{PB}(t) \times F_{PS}(t) \end{aligned} \tag{18}$$

We verify Equation (18) for generic probability CDF and PDF in HOL4 as:

Theorem 5.1.

```
⊢ ∀CS SS MA MS MB P B PA PB PS p t f_MA.
    0 ≤ t ∧ prob_space p ∧
    ALL_DISTINCT_RV [CS;SS;MA;MS;MB;P;B;PA;PB;PS] p t ∧
    indep_vars_sets [CS;SS;MA;MS;MB;P;B;PA;PB;PS] p t ∧
    distributed p lborel (λs.  real (MA s)) f_MA ∧ (∀y.  0 ≤ f_MA y) ∧
    cont_CDF p (λs.  real (MS s)) ∧
    measurable_CDF p (λs.  real (MS s)) ⇒
    (prob p
       (DFT_event p
           ((shared_spare PA PB PS PS)·(shared_spare PB PA PS PS)+
           (PAND MS MA)+(HSP MA MB)+
           (HSP (FDEP(CS + SS) P)(FDEP(CS + SS) B))) t) =
     CDF p (λs.  real (CS s)) t + CDF p (λs.  real (SS s)) t +
     pos_fn_integral lborel
        (λy.
           f_MA y * (indicator_fn {u | 0 ≤ u ∧ u ≤ t} y *
            CDF p (λs.  real (MS s)) y)) +
     CDF p (λs.  real (MA s)) t * CDF p (λs.  real (MB s)) t +
     CDF p (λs.  real (P s)) t * CDF p (λs.  real (B s)) t +
     CDF p (λs.  real (PA s)) t * CDF p (λs.  real (PB s)) t *
     CDF p (λs.  real (PS s)) t - ....+...-
     CDF p (λs.  real (CS s)) t * CDF p (λs.  real (SS s)) t *
     pos_fn_integral lborel
        (λy.
           f_MA y * (indicator_fn {u | 0 ≤ u ∧ u ≤ t} y *
            CDF p (λs.  real (MS s)) y)) *
     CDF p (λs.  real (MB s)) t * CDF p (λs.  real (P s)) t *
     CDF p (λs.  real (B s)) t * CDF p (λs.  real (PA s)) t *
     CDF p (λs.  real (PB s)) t * CDF p (λs.  real (PS s)) t)
```

where `0 ≤ t` ensures that the time t is greater than or equal to 0, `prob_space p` indicates that p is a probability space, `ALL_DISTINCT_RV` is a predicate which ensures that all inputs and their event sets are not equal and their values are greater than or equal to 0 but they cannot equal $+\infty$. This assumption is a realistic one, since for any component in a system the time of failure will always be greater than or equal to 0 and the component will eventually fail. The predicate `indep_vars_sets` adds the condition that all random variables and their event sets are independent. The predicate (`distributed p lborel (λs. real (MA s)) f_MA`) indicates that the real random variable of MA has the density function `f_MA`. The last two predicates in the goal ensures that the CDF of the real random variable of MS is continuous and measurable from the real line to the extreal one (real-borel to extreal-borel).

We verify several intermediate lemmas to prove Theorem 5.1. We first verify a reduced form of the given DFT and, then we verify the probability expression of the verified reduced version.

Lemma 5.1.

```
⊢ ∀CS SS MA MS MB P B PA PB PS.
    (∀s. ALL_DISTINCT [MA s; MS s; PA s; PB s; PS s]) ⇒
    ((shared_spare PA PB PS PS)·(shared_spare PB PA PS PS) +
     (PAND MS MA) +
     (HSP MA MB)+(HSP (FDEP(CS + SS) P)(FDEP(CS + SS) B)) =
     CS + SS + (MA·(MS ◁ MA)) + MA·MB + P·B + PA·PB·PS)
```

In the above lemma, `(shared_spare PA PB PS PS)·(shared_spare PB PA PS PS)` represents the pumps part of the DFT, `(PAND MS MA)+(HSP MA MB)` represents the motors parts and finally the CPUs part is represented by `(HSP (FDEP(CS + SS) P)(FDEP(CS + SS) B)`. The predicate `ALL_DISTINCT` ensures that all basic events cannot fail at the same time. Since we assumed that all spare gates are HSPs, the spare input PS for the shared spare gates is the same for both the active and dormant states.

In order to find the probability of the top event, we utilize the formally verified reduced version of the structure function and encapsulate it into a `DFT_event`, as the probability can only be applied to sets. To utilize the probabilistic PIE, we express the `DFT_event` of the Cardiac Assist system as the union of events.

Lemma 5.2.

```
⊢ ∀PA PB PS MS MA MB CS SS P B p t.
    DFT_event p
       (CS + SS + (MA·(MS ◁ MA)) + MA·MB + P·B + PA·PB·PS) t =
    union_list
     [DFT_event p CS t; DFT_event p SS t;
      DFT_event p (MA·(MS ◁ MA)) t;
      DFT_event p (MA·MB) t;
      DFT_event p (P·B) t; DFT_event p (PA·PB·PS) t]
```

From Lemma 5.2, we can notice that the top event is constructed from the union of six different sets. Applying the probabilistic PIE on the union of these sets (6 sets) generates 63 different terms (combinations). We verify several lemmas to be able to use the theorem of the probabilistic PIE [1] for the union list of these six sets. For example, we formally verify that:

Lemma 5.3.

```
⊢ ∀A B C D E K.
    {t | t SUBSET {A; B; C; D; E; k} ∧ t ≠ {}} =
    {{A}; {B}; {C}; {D}; {E}; {k}; {A; B}; {A; C};...;
     {A; B; C; D; E; k}}
```

The result of Lemma 5.3 is a set of 63 different sets. We had to apply the `SIGMA` function that results from the `sum_set` in the PIE theorem. Therefore we verify the following lemma for 63 sets.

Lemma 5.4.

```
⊢ ∀A B C D E K.
    ALL_DISTINCT [A;B;C;D;E;k] ∧
    (∀x.  x ∈{{A};{B};{C};{D};{E};{k};...;{A; B; C; D; E; k}} ⇒
       f x ≠ PosInf) ⇒
    (SIGMA f {{A};{B};...;{A; B; C; D; E; k}} =
     f {A} + f {B} +...+ f {A; B; C; D; E; k}
```

After verifying all these lemmas and based on the reduced DFT expression we are able to verify the probability of the Cardiac Assist system (Equation 18) into Theorem 5.1.

The first part of the conclusion of Theorem 5.1 corresponds to the original DFT (without reduction). In the verification of this theorem, we use Lemma 5.1 to replace the original DFT with the reduced one. Then, we use Lemma 5.2 to represent the `DFT_event` as a union list. After representing the left-hand side of the conclusion of Theorem 5.1 as a union list, we use Lemmas 5.3, 5.4 and the probabilistic PIE theorem [1] to prove this goal. After applying the probabilistic PIE, the resulting 63 subgoals should be proven based on the verified theorems of the probability of DFT gates. Therefore, applying the probabilistic PIE will not directly verify the current theorem, it is rather required to verify several intermediate subgoals after applying the PIE. In addition, after applying the PIE, it is necessary to apply the simplification theorems again since the application of the PIE results in intersecting the events. This means that further simplifications needed to be done. The first 6 terms in the right-hand side of the conclusion of Theorem 5.1 correspond to the probability of the elements of the list in Lemma 5.2. For example, `CDF p (λs. real (CS s)) t` represents the probability of `DFT_event p CS t`, which is $F_{CS}(t)$, according to Theorem 4.1. `pos_fn_integral lborel(λy. f_MA y *(indicator_fn {u |0 ≤ u ∧ u ≤ t} y * CDF p (λs. real (MS s)) y))` represents the probability of `DFT_event p (MA·(MS ◁ MA)) t`, which is $\int_0^t f_{MA}(y) \times F_{MS}(y)\, dy$, according to Theorem 4.4. The following terms in the conclusion of Theorem 5.1 correspond to finding the probability of the intersection of each pair in the list, then each 3 elements then 4 elements until we reach the last term in the right-hand side of the goal, which corresponds to the probability of the intersection of all elements in the list. Since all six elements in the union list are independent, the probability of their intersection is equal to the multiplication of the individual probabilities. The verification of Theorem 5.1 required around 6000 lines of proof script and took only 40 man-hours. This process was significantly facilitated thanks to the availability of the formally verified intermediate lemmas. The proof effort for the formal verification of these lemmas involved 12000 lines of code and about 320 man-hours. These lemmas can be easily reused for verifying similar systems that can be represented as the union of 6 events. In addition, these lemmas can also be useful in verifying larger systems based on similar proof steps identified above. As a future work, we plan to use machine learning techniques to automate the proof process of similar systems. This would facilitate the reusability of this work with users who are not much familiar with HOL or the underlying details of our verified theorems.

It is important to note that we have been able to verify the probability of the Cardiac Assist system for generic distributions and density functions, which can be instantiated later with specific functions according to the required constraints, without any need to repeat the whole process from the beginning. It is worth mentioning that such results cannot be obtained using PMCs, as they can only generate the probability of failure after specifying the failure rates of the components. In addition, PMCs are only limited to exponential distribution which does not consider the aging factor in any system. However, using our formalization for generic expression, it can be used with any probability distribution and density function as long as they are integrable, which makes it a more general and accurate alternative to the existing techniques.

6 Conclusions

In this paper, we proposed to conduct the probabilistic analysis of DFTs within the HOL4 theorem prover and thus obtain formally verified probability of failure expressions for generic probability distributions and density functions. We verified many simplification theorems for DFT gates and operators that allow formal reasoning about the reduction of the structure function of the DFT top event into a simpler form. In particular, we verified the probability of the intersection and the union of independent events to provide the probability of the AND, OR, FDEP and HSP gates. Moreover, we verified the probability of a sequence of two failing events ($Pr(X < Y)$) in two forms, i.e, $Pr(X < Y \wedge Y \leq t)$ and $Pr(X < Y \wedge X \leq t)$, which, to the best of our knowledge, is another novel contribution. These expressions are used to formally express the probability of the PAND gate and the before operator. Similarly, we also verified the probabilistic behavior of the spare gates, which required dealing with dependent events and conditional density functions. To illustrate the effectiveness of our formalization, we presented the formal analysis of the Cardiac Assist System, which is a safety-critical system. Using our formalization, we were able to provide generic results for the probability of failure of this system, i.e., for any distributions and density functions. It is evident that such results cannot be obtained using simulation nor using model checking. This highlights the importance of our proposed work, besides the fact that it inherits the sound and expressive nature of HOL theorem proving.

References

[1] W. Ahmad and O. Hasan. Towards Formal Fault Tree Analysis using Theorem Proving. In *Intelligent Computer Maths.*, LNCS 9150, pages 39–54. Springer, 2015.

[2] W. Ahmad and O. Hasan. Formalization of Fault Trees in Higher-order Logic: A Deep Embedding Approach. In *Dependable Software Engineering: Theories, Tools, and Applications*, LNCS 9984, pages 264–279. Springer, 2016.

[3] H. Bauer. *Probability Theory.* Walter de Gruyter, 1996.

[4] P. Billingsley. *Probability and Measure.* John Wiley & Sons, 2012.

[5] H. Boudali, P. Crouzen, and M. Stoelinga. A Rigorous, Compositional, and Extensible Framework for Dynamic Fault Tree Analysis. *IEEE Transactions on Dependable and Secure Computing*, 7:128–143, 2010.

[6] C. Dehnert, S. Junges, J.P. Katoen, and M. Volk. A Storm is Coming: A Modern Probabilistic Model Checker. In *Computer Aided Verification*, LNCS 10427, pages 592–600. Springer, 2017.

[7] Y. Elderhalli. DFT Formal Probabilistic Analysis: HOL4 Script, Concordia University, Montreal, QC, Canada, http://hvg.ece.concordia.ca/code/hol/DFT/index.php, 2018.

[8] Y. Elderhalli, O. Hasan, W. Ahmad, and S. Tahar. Formal Dynamic Fault Trees Analysis Using an Integration of Theorem Proving and Model Checking. In *NASA Formal Methods*, LNCS 10811, pages 139–156. Springer, 2018.

[9] M. Ghadhab, S. Junges, J.P. Katoen, M. Kuntz, and M. Volk. Model-based Safety Analysis for Vehicle Guidance Systems. In *Computer Safety, Reliability, and Security*, LNCS 10488, pages 3–19. Springer, 2017.

[10] HOL4. https://hol-theorem-prover.org/, 2018.

[11] J. Hölzl. *Construction and Stochastic Applications of Measure Spaces in Higher-Order Logic.* PhD thesis, Technische Universität München, Germany, 2012.

[12] G. Merle. *Algebraic Modelling of Dynamic Fault Trees, Contribution to Qualitative and Quantitative Analysis.* PhD thesis, ENS, France, 2010.

[13] T. Mhamdi. *Information-theoretic Analysis using Theorem Proving.* PhD thesis, Concordia University, Montreal, QC, Canada, 2012.

[14] T. Mhamdi, O. Hasan, and S. Tahar. On the Formalization of the Lebesgue Integration Theory in HOL. In *Interactive Theorem Proving*, LNCS 6172, pages 387–402. Springer, 2010.

[15] T. Mhamdi, O. Hasan, and S. Tahar. Formalization of Entropy Measures in HOL. In *Interactive Theorem Proving*, LNCS 6898, pages 233–248. Springer, 2011.

[16] T. Mhamdi, O. Hasan, and S. Tahar. Formalization of Measure Theory and Lebesgue Integration for Probabilistic Analysis in HOL. *ACM Transactions on Embedded Computing Systems*, 12(1):13, 2013.

[17] C.Z. Mooney. *Monte Carlo Simulation.* Sage, 1997.

[18] J. Ni, W. Tang, and Y. Xing. A Simple Algebra for Fault Tree Analysis of Static and Dynamic Systems. *IEEE Transactions on Reliability*, 62(4):846–861, 2013.

[19] L. Pullum and J.B. Dugan. Fault Tree Models for the Analysis of Complex Computer-based Systems. In *IEEE Reliability and Maintainability Symposium*, pages 200–207, 1996.

[20] M. Qasim. Formalization of Normal Random Variables. Master's thesis, Concordia University, Montreal, QC, Canada, 2016.

[21] M. Qasim, O. Hasan, M. Elleuch, and S. Tahar. Formalization of Normal Random Variables in HOL. In *Intelligent Computer Mathematics*, LNCS 9791, pages 44–59. Springer, 2016.

[22] E. Ruijters and M. Stoelinga. Fault Tree Analysis: A Survey of the State-of-the-art in Modeling, Analysis and Tools. *Computer Science Review*, 15-16:29 – 62, 2015.

[23] M. Stamatelatos, W. Vesely, J.B. Dugan, J. Fragola, J. Minarick, and J. Railsback. *Fault Tree Handbook with Aerospace Applications*. NASA Office of Safety and Mission Assurance, 2002.

[24] K. J. Sullivan, J. B. Dugan, and D. Coppit. The Galileo Fault Tree Analysis Tool. In *IEEE Symposium on Fault-Tolerant Computing*, pages 232–235, 1999.

Received 20 July 2018

A Novel Criterion for Rejecting the Non-Inductive Method

Ruurik Holm
Deduktia Ltd, Helsinki, Finland
ruurik.holm@deduktia.fi

Abstract

The article discusses the problem of choosing between inductive methods in Rudolf Carnap's inductive logic. It has been held that, if one has no background knowledge, there is no way to justify the use of one inductive method instead of another one. Even the non-inductive method, which gives no regard to experience, cannot be considered worse than the inductivist methods giving regard to experience. However, it can be shown that, by using Carnap's own success criterion for inductive methods, there is an inductivist method which is to be preferred over the non-inductive method. The result will be compared to Hans Reichenbach's pragmatic justification of induction.

1 Introduction

Rudolf Carnap defined a continuum of inductive methods in (1952), but did not establish convincing criteria on the basis of which one could choose the most appropriate method for each empirical situation. If no background knowledge is available, there seems to be no way to prefer any method to any other. In particular, there are no solid grounds for rejecting the non-inductive method, which yields the same probabilities regardless of any evidence.

It was shown by Good (1965) that Carnap's continuum is equivalent to Dirichlet's (prior) distributions in Bayesian statistics (see also Festa 1993, pp. 65-66; Festa 2011, pp. 477-478). Festa (1993; 1995; 2011, p. 485) discusses a possible procedure for obtaining the optimum inductive method. However, Festa relies on prior assumptions concerning the degree of order of the universe under scrutiny.

The "straight rule" projects observed frequencies of properties to estimations of the whole population. It will be shown that one should prefer the straight rule over the non-inductive method for all sample sizes. In a knowably non-homogeneous universe, this converts to a preference of a particular and knowable self-correcting method over the non-inductive method.

The argument is based on Carnap's criterion of the success of an inductive method, namely its mean square error. Other performance measures could also be presented, but being able to establish grounds for the straight rule for at least one measure of success gives at least one criterion for distinguisihing between inductive methods. Hence, to re-establish that it is not justified to resort to any inductivist method, one would at least need to show that the mean square error is not an acceptable criterion of success of inductive methods.

2 Formal Background

This article proceeds by following the presentation and terminology of Carnap (1952).

Carnap's continuum of inductive methods is defined for monadic predicate logic with the identity symbol $=$ between individual constants, which are finite in number and denoted by $a_1, ..., a_n$. (Although Carnap does not state this explicitly, it will be assumed that $a_i \neq a_j$ for all $i \neq j$.)

The molecular predicates are defined by using the usual logical connectives $\sim$, $\vee, \&$. (Carnap 1952, p. 9.)

The possible states of the universe are represented by state descriptions "containing as components for every atomic sentence either it or its negation, but not both, and no other sentences" (Carnap 1952, p. 11).

The inductive methods are conditional probability functions $c(h, e)$ (or in Carnap's terminology, methods of confirmation) or corresponding estimation functions (for Carnap, methods of estimation) defined in the usual way over the set of sentences (where the hypothesis h and evidence e are both sentences), and parametrised by the parameter λ, whose values are real numbers from 0 without an upper bound. Hence, the continuum of inductive methods is also referred to as the λ-continuum. The inductive method denoted by $\lambda = \infty$ is formally defined as the limit value of the method when $\lambda \to \infty$. (See Carnap 1952, pp. 32-33.)

The inductive methods of the λ-continuum fulfil certain conditions which are common for various systems of probability, denoted by $C1 - C10$ in Carnap (1952). For example, $C1$ states that if h and h' are logically equivalent sentences, their conditional probabilities (or "degree of confirmation", see e.g. Carnap 1952, p. 4) under evidence e are equal.

In Carnap's terminology, a *Q-predicate* is a conjunction which contains either the predicate itself or its negation for each (monadic) atomic predicate. A Q-predicate denotes the corresponding Q-property. Hence, a Q-predicate says of an individual to which it applies which atomic properties it possesses and which it does not. Each Q-predicate (Q-property) can thus be associated with a class of individual constants (individuals) of a particular type. The total set of Q-properties incorporates all possible combinations of atomic properties an individual can have.

Each molecular predicate M can be represented by a disjunction of Q-predicates. The logical width b of M is the number of disjuncts in such a disjunction (cf. Carnap 1952, p. 10).

The following formula in Carnap (1952, p. 33) gives the probability that the $x + 1$'th individual will satisfy the predicate M, given that the first x_M individuals satisfy it in the sample of size x:

$$\frac{x_M + \frac{b}{\kappa}\lambda}{x + \lambda}, \tag{1}$$

where κ is the number of Q-predicates in the language and b is the logical width of M. The same formula (1) expresses also the λ-estimate of the relative frequency $Est_{\lambda,\kappa}(x, M, w_n, x_M)$, where w_n is the state description of size n from which the sample is taken. (Cf. Carnap 1952, p. 33.)

Observe that the value of (1) does not depend on the size of the domain of individuals.

3 Self-Correcting Methods

Carnap himself regards the non-inductive method[1] $c^{\dagger}$, i.e. $\lambda = \infty$ as seemingly inappropriate for sound scientific reasoning on the grounds that it gives no consideration to experience when making expectations or estimations, if the experience does not concern the individual mentioned in the hypothesis (e.g., 1950 [1962], p. 564; 1952, p. 38). For example, the evidence of n black ravens does not affect the $c^{\dagger}$-probability that the $n+1$'th raven is black. This can be formally seen by taking the limit of (1) when $\lambda \to \infty$, which becomes $\frac{1}{\kappa}$ (Carnap 1952, p. 37).

What about the other λ-methods? Carnap puts forward the fact that all estimation methods based on the corresponding inductive methods are *self-correcting,* with the sole exception of $c^{\dagger}$ (1952, p. 63; also p. 44). To understand what this means, one has to consider Carnap's original treatment in more detail.

The original formulation of inductive logic in Carnap (1950, 1952) is based on finite domains of individuals (more precisely, individual constants) and thus also on

[1] By using the term "non-inductive method" I follow the terminology of Kuipers (1986, p. 38).

finite state descriptions. The probabilities in infinite domains are modelled by a limit procedure by letting n denoting the size of the domain tend to infinity.

In the discussion concerning the comparison of inductive methods, starting from page 56 in Carnap (1952), the size of the domain of individuals (and therefore the number of individuals occurring in the state descriptions) is assumed to be very large compared to the sample size x, i.e. $n >> x$. Carnap (1952, p. 20, pp. 62-63) introduces the class K_m $(m = n - x)$ of those m individuals described by the state description w_n not belonging to the sample w_x, $w_x \subset w_n$. In other words, the class K consists of individuals having particular properties as described by that part of w_n. Moreover, the elements of K are enumerated (Carnap 1952, p. 20). This means that one can speak about initial segments (author's terminology, not Carnap's) consisting of the first i elements of K.

The case of an infinite state description w_∞ is treated by Carnap by letting m in K_m grow without an upper bound. In other words, Carnap assumes that the infinite sequence K_∞ of individuals (and hence, also the infinite state description w_∞, conceived as a sequence of atomic sentences) is given somehow.

The extreme method $\lambda = 0$, which projects the observed frequencies to the whole domain, is called by Carnap the straight rule of confirmation (1952, p. 40) and the corresponding estimate function $Est_{0,\kappa}$ is called the straight rule of estimation. Carnap (1952, p. 63) refers to Reichenbach (1949), who shows that considering consecutive samples from w_∞ proves that his rule of induction (which is essentially the same as Carnap's $Est_{0,\kappa}$) will yield, for a given predicate, estimates approaching the real frequency of the predicate in w_∞, and is thus self-correcting.[2]

Carnap (1952, p. 63) proceeds to show that if $x \to \infty$ in (1), i.e. one takes samples of consecutive sizes (not necessarily consisting of initial segments of K_∞), then the value of (1) tends to the value of $Est_{0,\kappa}$ for a given positive finite λ, which, in turn, is a self-correcting method in the above sense when applied to consecutive samples (initial segments) from w_∞. Hence, all such methods are likewise self-correcting, i.e. if there is a limiting relative frequency r for a particular predicate, all methods except $c^\dagger$ converge toward r.

The self-correcting methods give better estimates than $c^\dagger$ only if the relative frequencies of all Q-predicates are not the same. One can say that a (finite) universe where each Q has the same relative frequency is maximally heterogeneous. It seems perhaps that a self-correcting method should be chosen because a maximally heterogeneous universe is such an extreme and unlikely case, but this is only an intuitive feeling based on our unconscious presupposition that the universe has some

[2] Actually, Carnap's elaboration takes place in terms of the infinite sequence K_∞, but this is slightly confusing since K_m is supposed to be the class of those individuals not mentioned in the given sample.

homogeneity.

However, there is a classical argument for choosing a self-correcting method analogous to the Reichenbachian pragmatic justification of induction (cf. Reichenbach 1949). This argument goes as follows: if there is a limiting relative frequency for a particular predicate, a self-correcting method will approach this frequency asymptotically, whereas the $c^{\dagger}$ method will not necessarily do so. In other words, if the universe is of the kind in which learning from experience is possible, then only the self-correcting methods will always approach the true relative frequencies, which is not the case with $c^{\dagger}$.

Unfortunately Reichenbach's justification of induction has a serious problem: convergence of the stream of data only means that the data converges to the real relative frequency *in the limit.* As Salmon (1991, p. 106) points out, "there is no finite integer N representing a sample size at which all of the regular asymptotic rules begin to converge". Hence, nothing guarantees or even makes it more probable that by using a self-correcting method, one actually obtains better estimates than by using a non-inductive method.

4 Carnap's Measure of Success: the Mean Square Error

Nevertheless, self-correcting methods can be argued for, but not in the traditional Reichenbachian way. The key is in the measure of success of inductive methods used by Carnap in (1952), the *mean square error,* which will be defined below.

It will be shown below that the mean square error of a particular self-correcting method, the straight rule, converges and has a computable maximum which is practically always smaller than that of the non-inductive method $c^{\dagger}$. Moreover, neither the maximum nor the real mean square error of $c^{\dagger}$ converge, which means that they cannot be made smaller by increasing the sample size.

The main difference of this line of argumentation to Reichenbach's justification is that one does not consider the actual sample from some very large state description – which may be anything, and thus the convergence to the real frequency can take place only in the limit. Instead one considers the average error of all possible samples of given size from that state description, in other words the square error of the average observer. In the absence of information about the best inductive method for the actual sample, it seams reasonable to choose the inductive method on the basis of its success on the average.

Carnap's notion of mean square error must now be introduced formally in order to derive the results mentioned above.

Carnap (1952, pp. 56-65) derives the mean square error for a given inductive method. Generally speaking, the mean square error is the mean of the square of the error of the estimate. If e denotes the estimate and y the actual value of a given magnitude, the error of the estimate is $e-y$. Since y is constant, the mean of $e-y$, i.e. $E(e-y)$, equals $E(e)-y$. With the help of variance of the error, $Var(e-y)$, one can write the mean square error as

$$E[(e-y)^2] = Var(e-y) + [E(e-y)]^2 \tag{2}$$

Let M be an arbitrary molecular predicate. When $n >> x$, the proportion of samples of size x with a given value of x_M (and hence the statistical probability that a random sample has a given value of x_M) can be approximated by

$$\binom{x}{x_M} r^{x_M}(1-r)^{x-x_M}, \tag{3}$$

where r is the relative frequency of M in the whole state description. Carnap (1952, pp. 62-63) shows that with the help of this and (2) above, one obtains the mean square error with respect to M.

However, that is not sufficient as a measure of success of an inductive method since it considers one Q-predicate only. Instead, Carnap (1952, pp. 65-67) defines as the measure of success the average mean square error of the estimates for all Q-predicates, which can be written as:

$$\frac{x - \frac{\lambda^2}{\kappa} + (\lambda^2 - x)\sum_{i=1}^{\kappa} r_i^2}{\kappa(x+\lambda)^2}, \tag{4}$$

Consider now the expression $\sum_{i=1}^{\kappa} r_i^2$ in (4) above, which is referred to as the *degree of order* of the whole state description (Carnap 1952, p. 66).

The degree of order represents how uniformly the individual constants have been distributed among the Q-predicates. For example, if all the individuals belong to the range of a single Q-predicate, the universe is extremely uniform (or homogeneous). In this case, the degree of order is 1. On the other hand, if no Q-predicate has more occurrences than any other Q-predicate, the universe is extremely non-uniform (or heterogeneous), in which case the degree of order is $\frac{1}{\kappa}$.

In the sequel, the degree of order $\sum_{i=1}^{\kappa} r_i^2$ will be denoted by d.

For every degree of order d, there is a corresponding optimum inductive method, which yields the smallest mean square error in all state desriptions whose degree of order is d:

$$\lambda(d) = \frac{1-d}{d-\frac{1}{\kappa}}. \tag{5}$$

4.1 Convergence and Maximum of the Mean Square Error

Let us proceed to analyse the convergence of the mean square error.

The expression (4) can be written as

$$\frac{x - \frac{\lambda^2}{\kappa} + \lambda^2 d - xd}{\kappa x^2 + 2\kappa\lambda x + \kappa\lambda^2}. \tag{6}$$

Suppose now that $\lambda = 0$. The equation (6) becomes

$$\frac{1-d}{\kappa x}. \tag{7}$$

If $d = 1$ (i.e., the degree of order of the actual state description is maximal), the value of (7) is zero. However, for all that is known of the actual state description, d can be as small as $\frac{1}{\kappa}$. For this value of d, the mean square error of the straight rule takes its maximum value:

$$\frac{1-\frac{1}{\kappa}}{\kappa x} = \frac{\frac{1}{\kappa} - \frac{1}{\kappa^2}}{x}. \tag{8}$$

For any number of primitive predicates, one knows exactly the maximum mean square error for each sample and, moreover, one can always make this maximum smaller by increasing the sample size. If the sample size is increased x-fold, the maximum mean square error reduces to $\frac{1}{x}$'th of the original value. For large samples, the maximum mean square error is very small compared to the sample of 1.

Let us now proceed to show that the straight rule yields almost always a smaller maximum mean square error than the non-inductive method.

We must thus examine the case when $\lambda = \infty$. The limit convention of Carnap (1952, p. 33) means that for any function $f(\lambda)$ the value of $f(\lambda)$, when $\lambda = \infty$, is $\lim_{\lambda\to\infty} f(\lambda)$. Hence, to calculate the mean square error of $\lambda = \infty$ for a particular x, one must consider the limit of (9) below when $\lambda \to \infty$. Beside (6), the mean square error (4) can also be written as

$$\begin{aligned} &\frac{x - \frac{\lambda^2}{\kappa} + \lambda^2 d - xd}{\kappa\lambda^2(\frac{x^2}{\lambda^2} + \frac{2x}{\lambda} + 1)} = \\ &\frac{\frac{x}{\kappa\lambda^2} - \frac{1}{\kappa^2} + \frac{d}{\kappa} - \frac{xd}{\kappa\lambda^2}}{\frac{x^2}{\lambda^2} + \frac{2x}{\lambda} + 1}. \end{aligned} \tag{9}$$

Consider the last form above when $\lambda \to \infty$. The first two terms in the denominator clearly tend to zero, which entails that the denominator tends to 1. The first and last terms of the numerator tend to zero as well. Hence, the whole expression tends to

$$\frac{d}{\kappa} - \frac{1}{\kappa^2} \tag{10}$$

for all values of x. This constant is the mean square error for $\lambda = \infty$. If $d = \frac{1}{\kappa}$, (10) is zero (as Carnap 1952, p. 69 also notes). However, if $d > \frac{1}{\kappa}$, (10) is greater than zero and does not converge. In particular, (10) reaches its maximum when $d = 1$,

$$\frac{1}{\kappa} - \frac{1}{\kappa^2}. \tag{11}$$

When (11) is compared to (8), the following can be stated:

Proposition 1. *The maximum mean square error of the straight rule is smaller than that of the non-inductive method for all $x > 1$, and for $x = 1$, the maximum errors are equal. Moreover, while (11) remains constant when the sample size (i.e., the value of x) increases, the function (8) decreases strictly toward zero.*

This shows that, for all samples sizes larger than one, the straight rule ($\lambda = 1$) is to be preferred over the non-inductive method ($\lambda = \infty$) when the maximum mean square error criterion is applied. However, the proof leaves open whether some finite values of λ should be preferred over $\lambda = 1$.

4.2 Straight Rule vs. Other Self-Correcting Methods

Carnap (1952, p 75-77) shows that for any evidence with at least two different kind of individuals, i.e. individuals manifesting two different Q-predicates, one can calculate a value $\lambda' > 0$ which is a lower bound for the optimum value of λ among those state descriptions which are compatible with the evidence. This value thus demonstrably yields a smaller mean square error than the straight rule $\lambda = 0$ for any state description which is compatible with the evidence.

However, in most cases this gives no practical guidance for choosing the optimum inductive method. While the mean square error with respect to all samples is smaller for λ' than for $\lambda = 0$, provided that the universe is non-homogeneous, i.e. $d < 1$, it is not shown by Carnap (1952) that this would hold for the narrower set of all non-homogeneous samples. Once we have a non-homogeneous piece of evidence, any larger sample consistent with this observation is also non-homogeneous. Hence, the average observer having the evidence that the state description is non-homogenous, is not necessarily better off by choosing λ' than any other method (since the set of such observers can only obtain non-homogeneous samples). This can be contrasted with the above result: the average observer is better off by choosing the straight rule instead of the non-inductive method, and the evidence obtained by the observer plays no role in determining which method one should use.

The only situation where Carnap's result is applicable is the case where it is somehow known without observation that $d < 1$. In such a case, the maximum mean square error of the straight rule is achieved when $d = \frac{1}{\kappa}$ and is given by (8) above. Since λ' yields now a smaller mean square error than the straight rule, the maximum mean square error of λ' is smaller than that of the straight rule – and thus also (almost always) smaller than the mean square error of the non-inductive method.

5 Philosophical Significance of the Result

Proposition 1 above certainly does not mean that for some particular sample, the error of the estimate of the straight rule (or the known method denoted by λ' in section 4.2) would be smaller that of the non-inductive method. However, by limiting the maximum average error (i.e., the maximum mean square error) when making estimates is an obvious reason for preferring the straight rule (or λ') instead of the non-inductive method.

Consider a situation where an estimate of the relative frequency is required. It is assumed that beside the sample, there is no information as regards the actual state description. Hence, one cannot choose the estimate on the basis of its *de facto* accuracy (if one could do this, no samples or estimates would be needed). It is natural to choose the method which yields the smallest expected value of the error when all possible state descriptions are considered. This is precisely the expected value of the square error of the estimate, i.e. the mean square error. The problem in this approach has been that there is no way to judge which inductive method gives the smallest mean square error. However, it was shown above that one can in fact know the maximum of the mean square error for the straight rule and it converges rapidly when the sample size increases. In other words, for a given sample, the estimate of the straight rule is, on the average, at least within a computable margin of the real relative frequency, and this margin can be made arbitrarily small by increasing the sample size.

This can be important information in a given decision-making situation. Hence, in the absence of additional information about the actual state description, the straight rule is a more rational basis for estimates and estimate-based decisions than the non-inductive method.

What is the significance of this result in terms of the problem of induction? It has been demonstrated above that in terms of estimating relative frequencies, it is justified to use the straight rule (or λ') instead of the non-inductive method. Using a self-correcting method amounts to making inductive inferences on the basis of evidence. This leads one to conclude that making inductive inferences is justified.

It can of course be objected that Carnap's assumption of a random sample from the unkown actual state description (cf. section 4 above) is not justified. What grounds does one have for assuming that the given sample is a random one and not biased? However, as Campbell & Franklin (2004) point out, unless there are specific reasons to think otherwise, it is justified to assume that a given sample is not biased.

The issue of the randomness assumption is too extensive to be discussed here in detail. But even if the random sample objection is valid, one has still provided an answer to the original problem of not being able to justify the use of a self-correcting method in inductive logic by using Carnap's measure of success of an inductive method.

References

[1] Campbell, S. & J.Franklin 2004: "Randomness and the justification of induction". *Synthese* 138, 79-99.

[2] Carnap, R. 1950 [1962]: *Logical Foundations of Probability,* (2nd edition in 1962, to which the page numbers refer). The University of Chicago Press, Chicago.

[3] Carnap, R. 1952: *The Continuum of Inductive Methods.* The University of Chicago Press, Chicago.

[4] Festa, R. 1993: *Optimum Inductive Methods.*Kluwer, Dordrecht.

[5] Festa, R. 1995: "Verisimilitude, Disorder, and Optimum Prior Probabilities". In T. Kuipers & R. Mackor (eds.), *Cognitive Patterns in Science and Common Sense.* Rodopi, Amsterdam.

[6] Festa, R. 2011: "Bayesian Inductive Logic, Verisimilitude, and Statistics". In p. S. Bandyopadhyay & M. R. Forster, *Philosophy of Statistics.* Elsevier, Oxford.

[7] Good, I. 1965: *The Estimation of Probabilities.* The MIT Press, Cambridge, Massachusetts.

[8] Kuipers, T. 1986: "Some estimates of the optimum inductive method". *Erkenntnis* 24, 37-46.

[9] Reichenbach, H. 1949: *The Theory of Probability.* University of California Press, Berkeley and Los Angeles.

[10] Salmon, W. 1991: 'Hans Reichenbach's Vindication of Induction". *Erkenntnis* 35, 99-122.

Received 8 January 2019

Adaptive Deontic Logics: a Survey

Frederik Van De Putte
Ghent University and University of Bayreuth
`frederik.vandeputte@ugent.be`

Mathieu Beirlaen
Ghent University
`mathieubeirlaen@gmail.com`

Joke Meheus
Ghent University
`joke.meheus@ugent.be`

Abstract

Adaptive Logics (ALs) are a viable and useful formal tool to handle various issues in deontic logic. In this paper, we motivate, explain, illustrate, and discuss the use of ALs in deontic logic. Published work on deontic ALs focusses mainly on conflicttolerant deontic logics (logics that can accommodate conflicting obligations) and – to a lesser extent – on problems concerning factual and deontic detachment. So does the present paper. Near the end of the paper, however, we also indicate some of the possibilities that the adaptive logic framework creates for tackling other types of problems within deontic logic.

Keywords: Deontic Logic, Adaptive Logics, Conflict-tolerance, Non-monotonic Reasoning, Benchmark Examples

We thank Jesse Heyninck, Stef Frijters, for valuable comments on previous drafts of the paper. We are also indebted to Lou Goble for his incisive remarks and useful suggestions. Finally, we thank Daniela Glavaničová for spotting some crucial errors in the penultimate version of this paper.

Frederik Van De Putte is a post-doctoral fellow of the Flemish Research Foundation (FWO-Vlaanderen) at Ghent University and a Marie Skłodowska-Curie Fellow at the University of Bayreuth. Mathieu Beirlaen was a post-doctoral fellow of the Flemish Research Foundation (FWO-Vlaanderen) at Ghent University at the time of writing this paper.

***Preludium*: Nathan's predicament**

One Friday evening, Nathan promises his mother that he will look after his little brother, Ben, on Saturday afternoon so that she can visit her sister. A couple of hours later, Nathan's girlfriend Lisa calls. Being a typical teenager and hopelessly in love, he completely forgets about the promise he made earlier to his mother and agrees with Lisa to go with her to the cinema on Saturday afternoon (to see this cool movie – children under the age of 13 not allowed!) and to go for a veggie burger in the evening. On Saturday, Lisa rings at the door. Almost simultaneously, his mother puts on her coat, meanwhile saying "So, I'll be back by five. Don't forget. . . ". Hearing this, Nathan remembers about *both* promises and immediately realizes what kind of situation he is in. Given his promises, there are several things he ought to do and it is clear that he cannot do them all. Keeping his promise to go for a veggie burger in the evening still seems feasible, but he cannot look after six year old Ben and at the same time take Lisa to this particular movie!

1 Introduction

Logical principles may fail to apply under certain conditions, and logical principles involving normative concepts are no exception. Even if we restrict our focus to the modalities "it is obligatory that" and "it is permitted that", there are circumstances in which we cannot apply certain plausible rules of inference (unrestrictedly) on pain of highly undesirable outcomes or even plain triviality.

The example from the *preludium* provides one kind of illustration of this phenomenon. It concerns a context in which an agent, in this case Nathan, faces several obligations that cannot be jointly fulfilled. In such contexts, several clusters of otherwise plausible principles involving obligations and permissions are problematic. Let us look at two instances of such clusters.

Consider first the combination of the principle that whatever is obligatory is also permissible (OIP), and the principle of the interdefinability of obligation and permission (ID):

(OIP) If A is obligatory, then A is also permitted: $\mathsf{O}A \supset \mathsf{P}A$
(ID) A is obligatory iff $\neg A$ is not permitted: $\mathsf{O}A \equiv \neg\mathsf{P}\neg A$

If both A and its negation $\neg A$ are obligatory ($\mathsf{O}A \wedge \mathsf{O}\neg A$), then by (ID) and the first conjunct, $\neg\mathsf{P}\neg A$. However, by (OIP) and the second conjunct, $\mathsf{P}\neg A$. So we obtain a plain contradiction: $\neg\mathsf{P}\neg A \wedge \mathsf{P}\neg A$. Even if one is willing to accept that contradictions are not absurd, it seems hard to accept that conflicting obligations entail them. Opinions may differ on which of these two principles is the most salient one. It is clear, however, that at least one of them has to be abandoned or adequately restricted if we want to avoid the outcome that conflicting obligations entail plain contradictions.

A second cluster of principles which is problematic in the face of conflicting obligations consists of the aggregation principle (Agg), the principle that "ought implies can" (OIC), and the impossibility of contradictory states of affairs (CP):

(Agg) If A and B are obligatory, then so is their conjunction: $(\mathsf{O}A \wedge \mathsf{O}B) \supset \mathsf{O}(A \wedge B)$
(OIC) If something is obligatory, then it is also possible: $\mathsf{O}A \supset \Diamond A$
(CP) Contradictions are impossible: $\neg\Diamond(A \wedge \neg A)$

If $\mathsf{O}A \wedge \mathsf{O}\neg A$, then, by (Agg), $\mathsf{O}(A \wedge \neg A)$ and hence by (OIC), $\Diamond(A \wedge \neg A)$. But this is in direct contradiction with (CP). Again, one of the principles from the cluster cannot be upheld (unrestrictedly) if we are to accommodate conflicting obligations, or at least if we want to avoid that such conflicts result in plain contradictions.

Besides conflicting obligations, there are other types of circumstances in which plausible logical principles may fail to apply. One that we want to consider here concerns the violation of conditional obligations, i.e. statements of the form "If A is the case, then B is obligatory" – formally, $\mathsf{O}(B \mid A)$. Each of the rules of factual detachment (FD) and deontic detachment (DD) is intuitively appealing as a rule for detaching unconditional obligations from conditional ones:

(FD) If it is obligatory that B given condition A, and if A is the case, then it is obligatory that B: $A, \mathsf{O}(B \mid A) \vdash \mathsf{O}B$
(DD) If it is obligatory that B given condition A, and if A is obligatory, then it is obligatory that B: $\mathsf{O}A, \mathsf{O}(B \mid A) \vdash \mathsf{O}B$.

The combination of (FD) and (DD) is known to cause trouble in so-called contrary-to-duty cases: cases in which a secondary obligation kicks in once a possibly conflicting primary obligation was violated. The following is an example of such a case.

Lisa and Nathan are a couple since eleven months. Lisa wants their first anniversary to be special and promises Nathan to take him to a "real" restaurant. One can only pay in cash at this restaurant, so if they are going to the restaurant, then Lisa ought to withdraw one hundred dollars at an ATM beforehand. However, on the day of the event, Lisa changes her mind and decides that she is not going to the restaurant after all – perhaps she is no longer sure she wants to be Nathan's girlfriend in the first place. In view of her promise, she (still) has the obligation to take Nathan to the restaurant: $\mathsf{O}A$. She also still has the conditional obligation that, if she takes Nathan there, she has to withdraw the money: $\mathsf{O}(B \mid A)$. However, if she is not going to any restaurant, then she should not withdraw a hundred dollars, since carrying around that much money for no reason would be hazardous: $\mathsf{O}(\neg B \mid \neg A)$. And as it happens to be, she is not going to the restaurant: $\neg A$.

Let us now see how the combination of (FD) and (DD) causes trouble for cases like this. If the obligation $\mathsf{O}A$ is violated, i.e. $\neg A$ is the case, then the primary conditional obligation $\mathsf{O}(B \mid A)$ leads to the unconditional obligation $\mathsf{O}B$ via (DD), while the secondary (contrary-to-duty) obligation $\mathsf{O}(\neg B \mid \neg A)$ leads to the unconditional obligation $\mathsf{O}\neg B$ via (FD). In order to resolve this conflict, we must block the application of (DD) or that of (FD).[1]

We will have much more to say about conflicting obligations and about the detachment of conditional obligations in the remainder of this paper. For now, these examples merely serve to illustrate a general point. In the circumstances described above – conflicting obligations and contrary-to-duty cases – one cannot consider principles such as the ones just mentioned as unrestrictedly valid. This leaves the logician who wants to explicate our reasoning in such cases with various options. One is to simply reject those principles, and hence declare a number of intuitive inferences simply invalid. Our stance towards this option is perhaps best summarized by the following words of van Benthem [96, p. 95]:

> This is like turning down the volume on your radio so as not to hear the bad news. You will not hear much good news either.

A more promising option is to look for restricted versions or alternative, more fine-grained formulations of those principles. For instance, for the case of conflicting obligations, one may argue that (Agg) should only be applicable in case the conjunction of A and B is possible. For contrary-to-duty cases, one may reformulate (FD) as a principle that concerns dynamic updates, rather than (mere) factual input – see e.g. [97] where this is proposed.

We will not pursue this second option here, even though occasionally we will show that some concrete instances of it fail to deliver an appropriate logic of normative reasoning, either on philosophical or on purely technical grounds. Instead, we will focus on a third option, i.e. to take (some of) these problematic principles to be only valid in a defeasible, context-sensitive way.

That this option seems well in line with our intuitions is easily demonstrated by returning to our examples. As soon as Nathan realizes that looking after Ben is

[1] Alternatively, we could bite the bullet and accept the outcome that both B and $\neg B$ are obligatory. But then our first illustration shows that we must give up other logical principles on pain of contradiction.

incompatible with going to that particular movie with Lisa, it seems quite rational to reject the conclusion that he ought to do both. But, suppose that his mother also made him promise to walk the family dog on Saturday evening. Would it be rational that, in view of the conflict concerning his afternoon plans, he also rejects the conclusion that he ought to go with Lisa for a veggie burger (at 6pm) and take the dog for a walk (at 10pm)? It seems that the one should have no bearing on the other. What this comes to is that, even if it makes sense to withdraw applications of (Agg) upon realizing that A and B are mutually exclusive, this need not affect other applications of (Agg).

In a similar vein, it seems quite natural that certain applications of (DD) are upheld *unless and until* it turns out that the unconditional obligation in the premises is violated. That Lisa has the obligation to withdraw money, even if she is not going to the restaurant at all, feels contra-intuitive to non-logicians. Is there something wrong with their intuitions? Not necessarily, and maybe even to the contrary. It seems quite justified that in cases like this, (DD) is treated as a *defeasible* rule of inference: the obligation is detached from the conditional obligation *provided* the unconditional obligation is not violated.

Note the difference between the third option and the first one. In our approach, we do not invalidate *principles*, we invalidate certain *applications* of principles and this is done only *when and where* necessary. This at once illustrates what we mean by context-sensitivity: whether an application of a certain principle or rule is validated or not depends on the specific context (the premises at issue).

The aforementioned clusters of principles governing obligations and permissions were originally introduced to hold unconditionally. The circumstances in which these principles are not (jointly) applicable, such as conflicts and violations, are often considered anomalous or exceptional. Other principles were acknowledged to be applicable only in a defeasible, context-sensitive manner right from their very introduction. We give only one example. Consider the *nullum crimen sine lege* principle: "If A is not forbidden, then A is permitted". This principle is best thought of as a kind of *default* rule: assume (or infer) $\mathsf{P}A$, unless $\mathsf{O}\neg A$ follows from the premises. This rule is defeasible by its very nature, in the sense that at least some of its instances are violated in every interesting application context.

In order to apply inference rules in a logic in a context-sensitive, defeasible manner, the consequence relation of this logic has to be *non-monotonic*: given a set of premises from which a conclusion A is derivable, it must be possible to revoke A in the light of additional premises.[2] *Adaptive logics* (henceforth, ALs) provide

[2] Formally, a logic $\mathbf{L}$ is non-monotonic iff (if and only if) there are two sets of formulas Γ and Δ and there is a formula A such that A is $\mathbf{L}$-derivable from Γ, while A is *not* $\mathbf{L}$-derivable from $\Gamma \cup \Delta$.

a natural way to explicate the premise-sensitive, defeasible application of certain inference rules in a formal logic.

ALs are built on top of a core logic, called the *lower limit logic*, the inference rules of which hold unconditionally and unrestrictedly. An AL strengthens its lower limit logic by allowing a number of additional inference rules to be applied relative to the specific premises at hand. The term "adaptive logic" originates from this premise-sensitivity: ALs "adapt" themselves to the premises under consideration.

Beside ALs, many other formalisms for modelling defeasible reasoning have been applied in a deontic context: default logic [53], defeasible deontic logic [71], formal argumentation theories [33; 77; 92; 18; 105], input/output logic [76], etc. These different frameworks are all linked to one another and to ALs in various ways – see e.g. [47] for some recent comparisons.

There is, however, a distinctive feature of ALs that sets them apart from other approaches to non-monotonic reasoning, viz. their dynamic proof theory. The idea behind this proof theory is that the non-monotonicity of the logic's consequence relation is pushed into the object-level proofs. This means that a given derivation in a proof can become rejected in the light of other derivations within that same proof.[3]

Another important difference between the existing work on ALs and other types of non-monotonic logics is the pivotal role that classical logic (henceforth **CL**) plays within the latter. ALs are, at least in origin, more pluralistic in spirit regarding the meaning of the classical connectives, thus opening up to new perspectives on defeasible reasoning that are hard to detect when one sticks to **CL** as one's underlying monotonic logic.[4]

The current paper's aim is to motivate, explain, illustrate, and discuss the use of ALs in deontic logic. Published work on deontic ALs focusses mainly on conflict-tolerant deontic logics (logics that can accommodate conflicting obligations) and – to a lesser extent – on problems concerning factual and deontic detachment. So does the present paper. Near the end of the paper, however, we also indicate some of the possibilities that the adaptive logic framework creates for tackling other types of problems within deontic logic.

The outline of this paper is as follows. For ease of reference, we start by recalling the basic definitions concerning Standard Deontic logic, henceforth **SDL** (Section 2). In Section 3 we provide an introduction to the framework of ALs. By way of illustration, we first present two very simple adaptive logics that can handle examples as the one from the preludium (Section 3.1).

[3] We will define and illustrate the dynamic proof theory of ALs in Section 3.

[4] This aspect of ALs is nicely illustrated by our Section 7, where we introduce and discuss (adaptive) paraconsistent deontic logics.

In Sections 5–7 we present and discuss a variety of conflict-tolerant deontic ALs that move further away from the standard view: unlike the logics from Section 3.1, the logics from Sections 5–7 have lower limit logics that are inferentially weaker than **SDL**. Section 4 provides the conceptual and technical basis for this discussion. Whereas Sections 5 and 6 are mainly based on existing work, Section 7 presents mostly new material that we think improves on the existing work in a number of ways – we explain this in Section 7.4.

Section 8 summarizes the merits and demerits of the conflict-tolerant ALs presented throughout Sections 3–7. In that section we also show how the simple logics introduced in Section 3.1 can be further refined in various ways.

The other main application of existing deontic ALs concerns the problem of detaching conditional obligations. We distinguish between various approaches to this problem in Section 9, and discuss adaptive versions of each of them.

In Section 10 we show how the *nullum crimen sine lege* principle can be captured within the AL framework, and how this gives rise to various extensions of the logics defined in previous sections. This at once paves the way for our last section in which we give a short summary of the paper and point to ideas for future research.

Throughout this paper our focus is on the illustration and motivation of the core ideas we present, rather than on formal details and meta-theoretical results. Whenever relevant, we provide pointers to the literature, cf. the subsections "further reading and open ends".

Much of what we will write in this paper builds on Lou Goble's work on normative conflicts, which is nicely summarized in [42]. We will provide references to specific parts of this (and other) work in due course. In general, we try to avoid overlap as much as possible, but whenever this maxim conflicts with keeping the present paper self-contained, we give priority to the latter.

We end this section with some more general comments regarding the plurality and diversity of logics to be discussed in this paper. Our stance on the matter can be described as follows.

For a start, various logics present themselves as useful depending on the specific type of application context, and the associated logical grammar one wants to study. But even if we keep the grammar fixed, there are various reasons for occupying oneself with not one but many logics for this grammar. That logic – even the logic of our most basic connectives like conjunction – is not god-given, and that there are no absolute grounds for preferring one logic over another, seems hardly contested nowadays. So all one can do is give pragmatic arguments, referring to general desiderata for logics on the one hand, and the needs of a given application on the other.

In the context of conflict-tolerant deontic logics, one way to argue for diversity is by referring to various explosion principles, as discussed in Section 4.2. For instance, if one does not *need* to accommodate conflicts between obligations and permissions, or if one can safely assume within a given domain that norms are at least internally consistent, then this should translate to one's preferred logic for that domain. Moreover, there are many different ways one can interpret the O of a given (conflict-tolerant or other) deontic logic, which will yield different formal semantics and hence different logics in turn.

Going non-monotonic (or in our case, going adaptive) does not reduce this plurality – quite to the contrary. To use Makinson's words [61, p. 14]:

> Leaving technical details aside, the essential message is as follows. Don't expect to find *the* nonmonotonic consequence relation that will always, in all contexts, be the right one to use. Rather, expect to find several *families* of such relations, interesting *syntactic conditions* that they sometimes satisfy but sometimes fail, and principal *ways of generating* them mathematically from underlying structures.

Indeed, it will become clear throughout this paper that there are usually several interesting and sensible ways of going adaptive, starting from a given lower limit logic. In the absence of further philosophical arguments against the resulting logics, one needs to keep an open mind and study all of them.

2 Some formal preliminaries

Languages Throughout this paper, we use $A, B, \ldots$ as metavariables for formulas of a given formal language, and $\Gamma, \Delta, \ldots$ as metavariables for sets of such formulas.

Let henceforth **CL** stand for the propositional fragment of classical logic, as based on a set of propositional variables (also called sentential letters) $\mathcal{S} = \{p, q, \ldots\}$, the connectives $\neg, \vee, \wedge, \supset, \equiv$, and the logical constants $\bot, \top$. We use $\mathcal{W}$ to denote the set of well-formed formulas in this language.

The language of **SDL** is obtained by adding to the grammar of **CL** the modal operators O for "it is obligatory that" and P for "it is permitted that". We take both O and P (and the classical connectives) to be primitive by default in this paper; i.e. whenever one is defined in terms of the others in one logic or another, we will indicate so. For the sake of simplicity, we will focus on the fragment of this language in which no nested occurrences of O and P are allowed. This means that the set of well-formed formulas for **SDL** is defined as follows:

$$\begin{aligned}\mathcal{W}^d := \quad & \mathcal{W} \mid \neg\langle\mathcal{W}^d\rangle \mid \langle\mathcal{W}^d\rangle \vee \langle\mathcal{W}^d\rangle \mid \langle\mathcal{W}^d\rangle \wedge \langle\mathcal{W}^d\rangle \mid \langle\mathcal{W}^d\rangle \supset \langle\mathcal{W}^d\rangle \mid \\ & \langle\mathcal{W}^d\rangle \equiv \langle\mathcal{W}^d\rangle \mid \mathsf{O}\langle\mathcal{W}\rangle \mid \mathsf{P}\langle\mathcal{W}\rangle\end{aligned}$$

Axiomatization The logic **SDL** is obtained by adding to **CL** the following axioms, rule, and definition:

(K) $\mathsf{O}(A \supset B) \supset (\mathsf{O}A \supset \mathsf{O}B)$
(D) $\mathsf{O}A \supset \neg\mathsf{O}\neg A$
(N) if $\vdash A$, then $\vdash \mathsf{O}A$
(Def$_\mathsf{P}$) $\mathsf{P}A =_{\mathsf{df}} \neg\mathsf{O}\neg A$

It is well-known that in the presence of (N), (K) can equivalently be expressed as the combination of the axiom of aggregation (Agg) and the rule of inheritance (Inh):

(Agg) $(\mathsf{O}A \wedge \mathsf{O}B) \supset \mathsf{O}(A \wedge B)$
(Inh) if $\vdash A \supset B$, then $\vdash \mathsf{O}A \supset \mathsf{O}B$

whence **SDL** can be equivalently characterized by adding (N), (Agg), (Inh), (D), and (Def$_\mathsf{P}$) to **CL**. Note also that in the presence of (Agg), (D) is equivalent to the following principle:

(P) $\neg\mathsf{O}(A \wedge \neg A)$

For ease of reference, we note some more derivable principles of **SDL**. The first is the axiom of distributivity (of O over $\wedge$):

(Dist) $\mathsf{O}(A \wedge B) \supset (\mathsf{O}A \wedge \mathsf{O}B)$

Second, the replacement of equivalents rule (RE) is an immediate consequence of the behavior of $\supset$ and $\equiv$ in **CL** and (Inh):

(RE) if $\vdash A \equiv B$, then $\vdash \mathsf{O}A \equiv \mathsf{O}B$

Third and last, in view of (Agg), (Inh), and the validity of disjunctive syllogism (DS) in **CL**, we have:

(DDS) $(\mathsf{O}A \wedge \mathsf{O}(\neg A \vee B)) \supset \mathsf{O}B$

Semantics We work with the traditional Kripke-semantics for **SDL**, but to prepare for the semantics of other logics to be presented below, we work with a designated "actual" world. An **SDL**-model M is a quadruple $\langle W, w_0, R, v\rangle$, where W is a non-empty set of worlds, $w_0 \in W$ is the actual world, $R \subseteq W \times W$ is a serial[5] accessibility relation and $v : W \to \mathcal{S}$ is a valuation function. $R(w)$ (the image

[5] R is serial iff for every $w \in W$, there is a $w' \in W$ such that $(w, w') \in R$.

of w under R) is the set of worlds that are accessible from the viewpoint of w, $R(w) = \{w' \mid (w, w') \in R\}$.

The semantic clauses for the sentential variables and the connectives are as usual; those for O and P are as follows:

(SC1) $M, w \models \mathsf{O}A$ iff $M, w' \models A$ for all $w' \in R(w)$
(SC2) $M, w \models \mathsf{P}A$ iff $M, w' \models A$ for some $w' \in R(w)$

Truth of a formula A at a world w is given by the relation $\models$. Truth in a model $M = \langle W, w_0, R, v\rangle$ is simply truth at w_0. We say that M is a model of Γ iff all the members of Γ are true in M, i.e. iff for all $B \in \Gamma$, $M, w_0 \models B$. Semantic consequence is then defined as the preservation of truth in all models: $\Gamma \Vdash A$ iff A is true in all models of Γ.

Following customary notation, let $|A|_M =_{\mathsf{df}} \{w \mid M, w \models A\}$. $|A|_M$ is also called the *truth set* (intension) of A. Note that the semantic clause for O can be equivalently rewritten as follows: $M, w \models \mathsf{O}A$ iff $R(w) \subseteq |A|_M$.

3 Adaptive logics

Adaptive logics were originally introduced by Diderik Batens around the 1980s, and have since been applied to various forms of defeasible reasoning.[6] The aim of this section is to highlight the basic features of ALs by means of a running example, viz. the logics $\mathbf{SDL}^{\mathbf{r}}_{\mathbf{p}}$ and $\mathbf{SDL}^{\mathbf{m}}_{\mathbf{p}}$. These logics can handle simple cases of conflicting obligations such as the running example from the beginning of this paper. We explain the idea behind both logics in Section 3.1. Generic definitions for all ALs in the standard format from [10] are given in Section 3.2. We mention the most salient properties of all logics that are defined within this format in Section 3.3. Finally, we discuss some variants of the standard format that will turn out useful in the remainder of this paper (Section 3.4).

3.1 The basics

Before introducing the logics $\mathbf{SDL}^{\mathbf{r}}_{\mathbf{p}}$ and $\mathbf{SDL}^{\mathbf{m}}_{\mathbf{p}}$, we present another predicament from Nathan's life. The example will be used to illustrate the proof theory of $\mathbf{SDL}^{\mathbf{r}}_{\mathbf{p}}$ and $\mathbf{SDL}^{\mathbf{m}}_{\mathbf{p}}$.

One evening, Nathan comes home from school. As soon as he enters the kitchen, he hears his father: "Remember, Nathan, it's your turn to do the dishes tonight. Do them this time!" His mother immediately adds: "And forget about playing with

[6] See Section 3.5 for references to the literature on ALs.

Ben tonight. Before supper, you will do nothing but your homework. Your grades are terrible lately!" Not too enthusiastically, Nathan heads towards his room to do his homework. As soon as he wants to enter it, his twin sister Olivia leaves hers, in great despair: "Nathan, you have to help me. I am on "Ben watch" tonight, but he is driving me crazy and I am expecting this really, really important phone call! Play with him until supper, will you? I'll do anything for you in return!" Nathan finds himself again in a difficult situation. He can obey his father and do the dishes. No problem there. But what should he do until supper? Olivia helps him out on a quite regular basis and he feels he ought to return the favor this time. But if he plays with Ben, he will not be able to do his homework.

This example and the one from the *preludium* have three important characteristics in common. The first is that they both concern a situation in which an agent faces several obligations, not all of which can be fulfilled. The second is that, for each of the separate obligations, there is some *prima facie* reason. In the example from the *preludium*, Nathan's specific obligations hold in view of the general rule "One ought to keep one's promises". In this last example, the obligation that Nathan ought to do the dishes holds in view of his father's command. The third characteristic is that, although not all obligations can be met, some of them can. Nathan cannot look after Ben and take Lisa to that particular movie, but he *can* go for a veggie burger in the evening. Similarly, Nathan cannot do his homework and at the same time play with Ben, but he *can* do the dishes.

In this paper, we will use the term *prima facie obligations* for any obligation for which there is some *prima facie* reason (some general rule, a command, . . .). As the examples show (and as we all know from daily life), there are situations in which not all *prima facie* obligations can be *binding*. Nathan cannot go to that particular movie with Lisa (in view of his promise to her) and at the same time *not* go there (in view of his promise to his mother and the fact that six year olds are not allowed for this particular movie). We will use the term *actual obligations* for obligations that are binding and that should be acted upon.

Examples in which not all *prima facie* obligations can be met raise the following question: how do we decide, in a given situation, which *prima facie* obligations are actual obligations and which are not? A first answer to this question seems to be that at least those *prima facie* obligations should be considered as actual obligations that are not in conflict with any other *prima facie* obligation. This seems to capture nicely our intuitions behind the examples. The fact that Nathan made conflicting promises with respect to what he will do in the afternoon should not prevent him from going for a veggie burger in the evening. The fact that he cannot help out his twin sister as well as obey his mother should not rule out that he at least obeys his father.

This is exactly the idea behind the logics $\mathbf{SDL^r_p}$ and $\mathbf{SDL^m_p}$ presented in this section: *prima facie* obligations are considered as actual obligations *unless and until* it turns out that they are in conflict with some other *prima facie* obligation. Or, put in a somewhat different form, the logics $\mathbf{SDL^r_p}$ and $\mathbf{SDL^m_p}$ validate the inference of actual obligations from *prima facie* obligations *as much as possible.* The exact meaning of this "as much as possible" will become clear below.

The logics have two further characteristics in common: they allow us to (a) accommodate conflicts at the level of *prima facie* obligations, and (b) reason about actual obligations in the standard way (i.e., applying all axioms of **SDL**).[7] What (a) comes to is that both logics are conflict-tolerant: they do not lead to unwanted conclusions in the face of conflicting *prima facie* obligations.

We will now show, step by step, how the logics $\mathbf{SDL^r_p}$ and $\mathbf{SDL^m_p}$ are obtained.

The lower limit logic In order to make the distinction between *prima facie* obligations and actual obligations, we will use a bi-modal language that contains two obligation operators: $\mathsf{O^p}$ and O. The first is used for *prima facie* obligations, the second for actual obligations. The language is defined as follows:

$$\begin{aligned}\mathcal{W}^{\mathsf{p}} := \quad & \mathcal{W} \mid \mathsf{O}\langle\mathcal{W}\rangle \mid \mathsf{O^p}\langle\mathcal{W}\rangle \mid \neg\langle\mathcal{W}^{\mathsf{p}}\rangle \mid \langle\mathcal{W}^{\mathsf{p}}\rangle \vee \langle\mathcal{W}^{\mathsf{p}}\rangle \mid \langle\mathcal{W}^{\mathsf{p}}\rangle \supset \langle\mathcal{W}^{\mathsf{p}}\rangle \mid \\ & \langle\mathcal{W}^{\mathsf{p}}\rangle \wedge \langle\mathcal{W}^{\mathsf{p}}\rangle \mid \langle\mathcal{W}^{\mathsf{p}}\rangle \equiv \langle\mathcal{W}^{\mathsf{p}}\rangle\end{aligned}$$

Note that we exclude nesting; i.e. none of the two operators occurs within the scope of another operator.

To obtain a logic that is tolerant with respect to conflicting *prima facie* obligations (characteristic (a) above), $\mathsf{O^p}$ is treated as a propertyless operator, a "dummy". This means that e.g. *prima facie* obligations cannot be derived from other *prima facie* obligations. Characteristic (b) is realized by assuming that O is the ought-operator of **SDL**.

Let us call the resulting logic $\mathbf{SDL_p}$ – it is just **SDL** extended with the dummy-operator $\mathsf{O^p}$. In AL terminology, what we have done so far is define the *lower limit logic* of our AL. This logic constitutes the monotonic core of the AL. In other words, it consists of all the principles (rules, axioms) that are unconditionally valid within the logic.[8]

In order to obtain a logic that validates the inference from *prima facie* obligations to actual obligations as much as possible, $\mathbf{SDL_p}$ needs to be strengthened. One option that does *not* work is to simply add the axiom

[7] Our characteristics (a) and (b) correspond to Goble's criteria of adequacy a) and b) for *prima facie* oughts versus all-things-considered oughts [42, p. 257].

[8] The lower limit logic of every AL has to satisfy certain general desiderata, which will be spelled out in Section 3.2.

(A) $\mathsf{O}^{\mathsf{p}}A \supset \mathsf{O}A$

to $\mathbf{SDL_p}$. Let us call the resulting logic $\mathbf{SDL_p^+}$. In this stronger logic, conflicts at the level of *prima facie* obligations will be trivialized: if $\vdash_{\mathbf{CL}} \neg(A_1 \wedge \ldots \wedge A_n)$, then $\mathsf{O}^{\mathsf{p}}A_1, \ldots, \mathsf{O}^{\mathsf{p}}A_n \vdash_{\mathbf{SDL_p^+}} B$ for any B.[9] Of course, we could weaken the logic of O, but then we would lose characteristic (b). This shows that we need a more refined way to fulfill our aim. We will now show how this can be realized within the framework of adaptive logics.

Going adaptive What we need is a way to steer between $\mathbf{SDL_p}$ and $\mathbf{SDL_p^+}$, avoiding the weakness of the former but also the explosive character of the latter. More precisely, we need a defeasible, context-sensitive version of (A). This can be done by assuming that formulas like $\mathsf{O}^{\mathsf{p}}p \wedge \neg\mathsf{O}p$, $\mathsf{O}^{\mathsf{p}}q \wedge \neg\mathsf{O}q$, etc. are false *unless and until* proven otherwise.

In AL terminology, such formulas – the negations of defeasible assumptions – are called *abnormalities*.[10] We will use Ω_{p} to refer to the set of all those abnormalities, i.e. all formulas of the form $\mathsf{O}^{\mathsf{p}}A \wedge \neg\mathsf{O}A$.

In an adaptive proof, we can derive formulas on the assumption that certain abnormalities are false. This is most easily illustrated with an example. Let d stand for "Nathan washes the dishes", b for "Nathan plays with Ben" and h for "Nathan does his homework". The *prima facie* obligations that Nathan faces in our second running example may then be formalized as $\mathsf{O}^{\mathsf{p}}d$, $\mathsf{O}^{\mathsf{p}}b$ and $\mathsf{O}^{\mathsf{p}}(\neg b \wedge h)$. An adaptive proof from $\Gamma = \{\mathsf{O}^{\mathsf{p}}d, \mathsf{O}^{\mathsf{p}}b, \mathsf{O}^{\mathsf{p}}(\neg b \wedge h)\}$ in which we try to derive the actual obligation for Nathan to wash the dishes ($\mathsf{O}d$) may then look as follows:

1	$\mathsf{O}^{\mathsf{p}}d$	Prem	$\emptyset$
2	$\mathsf{O}^{\mathsf{p}}b$	Prem	$\emptyset$
3	$\mathsf{O}^{\mathsf{p}}(\neg b \wedge h)$	Prem	$\emptyset$
4	$\mathsf{O}d \vee \neg\mathsf{O}d$	SDL	$\emptyset$
5	$\mathsf{O}d \vee (\mathsf{O}^{\mathsf{p}}d \wedge \neg\mathsf{O}d)$	1,4; SDL	$\emptyset$
6	$\mathsf{O}d$	5; RC	$\{\mathsf{O}^{\mathsf{p}}d \wedge \neg\mathsf{O}d\}$

The fourth element of each line in this proof represents the condition of that line. This condition is always a (possibly empty) set of abnormalities. After introducing

[9] To see why, note that in $\mathbf{SDL_p}$, conflicting actual obligations are trivialized just as in **SDL**. If we moreover allow for the unrestricted application of (A), this means that also conflicts at the level of *prima facie* obligations are trivialized.

[10] Our terminology here and below suggests a link with Makinson's *Default Assumption Consequence Relations* [61]. Indeed, as shown in [99], one can establish an exact correspondence between Makinson's construction and ALs that use the minimal abnormality strategy.

the premises on lines 1-3, we have used excluded middle to derive a new formula at line 4, and then derived line 5 using lines 1 and 4. We use "SDL" as a generic name for all rules and axioms of **SDL**. At line 6, $\mathsf{O}d$ is derived on the condition that the abnormality $\mathsf{O}^{\mathsf{p}}d \wedge \neg\mathsf{O}d$ is false. This is done by means of the rule RC (shorthand for *conditional rule*) which allows us to push abnormalities to the condition within an adaptive proof.

Here are two other applications of RC:

$\vdots$	$\vdots$	$\vdots$	$\vdots$
7	$\mathsf{O}b \vee (\mathsf{O}^{\mathsf{p}}b \wedge \neg\mathsf{O}b)$	2; SDL	$\emptyset$
8	$\mathsf{O}b$	7; RC	$\{\mathsf{O}^{\mathsf{p}}b \wedge \neg\mathsf{O}b\}$
9	$\mathsf{O}(\neg b \wedge h) \vee (\mathsf{O}^{\mathsf{p}}(\neg b \wedge h) \wedge \neg\mathsf{O}(\neg b \wedge h))$	3; SDL	$\emptyset$
10	$\mathsf{O}(\neg b \wedge h)$	9; RC	$\{\mathsf{O}^{\mathsf{p}}(\neg b \wedge h) \wedge$ $\neg\mathsf{O}(\neg b \wedge h)\}$

At this point, the reader may become suspicious. Clearly, $\mathsf{O}b$ and $\mathsf{O}(\neg b \wedge h)$ cannot both be true. By means of well-known **SDL**-principles, we can derive from our premises that at least one of the two corresponding abnormalities is true:

11	$(\mathsf{O}^{\mathsf{p}}b \wedge \neg\mathsf{O}b) \vee (\mathsf{O}^{\mathsf{p}}(\neg b \wedge h) \wedge \neg\mathsf{O}(\neg b \wedge h))$	2,3; SDL	$\emptyset$

Formulas like the one at line 11 are called Dab-*formulas* (Dab is shorthand for "disjunction of abnormalities"). Note that this Dab-formula is derived on the empty condition. Hence, it is an unconditional consequence of the premises – it cannot be false, if the premises are true. Moreover, it is *minimal*: neither of its disjuncts $\mathsf{O}^{\mathsf{p}}b \wedge \neg\mathsf{O}b$ or $\mathsf{O}^{\mathsf{p}}(\neg b \wedge h) \wedge \neg\mathsf{O}(\neg b \wedge h)$ is derived on the empty condition in the above proof.[11]

At lines 8 and 10 respectively, we relied on the assumption that the first, respectively the second of these abnormalities is false. But line 11 clearly indicates that those two assumptions cannot be jointly true. So a mechanism is needed to *retract* the inferences at lines 8 and 10.

Formally, this is taken care of by a *marking definition*, which stipulates which lines are marked, and hence considered "out" at a given stage of an adaptive proof. How the marking proceeds depends on the so-called *adaptive strategy*. The logics $\mathbf{SDL}^{\mathbf{r}}_{\mathbf{p}}$ and $\mathbf{SDL}^{\mathbf{m}}_{\mathbf{p}}$ are based respectively based on the Reliability strategy and the Minimal Abnormality strategy. Let us look at these in turn.

[11] In fact, neither of them *can* be derived in this proof on the empty condition, since they simply do not follow from Γ by $\mathbf{SDL}_{\mathbf{p}}$.

Reliability For $\mathbf{SDL}^{\mathbf{r}}_{\mathbf{p}}$, a line is marked whenever its condition contains an abnormality that is a disjunct of a minimal Dab-formula that has been derived in the same proof. For instance, in the above example, lines 8 and 10 are marked, whereas all other lines are not marked. This is indicated by a ✓-symbol at the end of the line:

1	$\mathsf{O}^{\mathsf{p}}d$	Prem	$\emptyset$
2	$\mathsf{O}^{\mathsf{p}}b$	Prem	$\emptyset$
3	$\mathsf{O}^{\mathsf{p}}(\neg b \wedge h)$	Prem	$\emptyset$
4	$\mathsf{O}d \vee \neg\mathsf{O}d$	SDL	$\emptyset$
5	$\mathsf{O}d \vee (\mathsf{O}^{\mathsf{p}}d \wedge \neg\mathsf{O}d)$	1, 4; SDL	$\emptyset$
6	$\mathsf{O}d$	5; RC	$\{\mathsf{O}^{\mathsf{p}}d \wedge \neg\mathsf{O}d\}$
7	$\mathsf{O}b \vee (\mathsf{O}^{\mathsf{p}}b \wedge \neg\mathsf{O}b)$	2; SDL	$\emptyset$
8	$\mathsf{O}b$	7; RC	$\{\mathsf{O}^{\mathsf{p}}b \wedge \neg\mathsf{O}b\}$ ✓
9	$\mathsf{O}(\neg b \wedge h) \vee (\mathsf{O}^{\mathsf{p}}(\neg b \wedge h) \wedge \neg\mathsf{O}(\neg b \wedge h))$	3; SDL	$\emptyset$
10	$\mathsf{O}(\neg b \wedge h)$	9; RC	$\{\mathsf{O}^{\mathsf{p}}(\neg b \wedge h) \wedge \neg\mathsf{O}(\neg b \wedge h)\}$ ✓
11	$(\mathsf{O}^{\mathsf{p}}b \wedge \neg\mathsf{O}b) \vee (\mathsf{O}^{\mathsf{p}}(\neg b \wedge h) \wedge \neg\mathsf{O}(\neg b \wedge h))$	2,3; SDL	$\emptyset$

In general, lines with an empty condition are never marked. But also those lines whose condition is not problematic in view of the minimal Dab-formulas in the proof remain unmarked (witness line 6 in the example). So at the end of the day, some instances of (A) are trustworthy in the light of the premises, while other instances of (A) are not. This illustrates the premise-sensitivity of adaptive logics that was mentioned in Section 1.[12]

The fact that lines can become marked in a proof means that we cannot simply define logical consequence in terms of being derivable in a proof. We need a more robust notion of derivability; this is called *final derivability.* The basic idea is that something is finally derivable if and only if it can be derived in a "stable" way. Spelling out this intuition is not as straightforward as it may seem, as it requires quantification over extensions of proofs. We refer to Definitions 3.3 and 3.4 in the next section for the exact details.

[12] Some may argue that, in light of the premise set, the inferences at lines 8 and 10 were never rational in the first place. Admittedly, in cases like Γ above, it can easily be seen which *prima facie* obligations can make it into actual obligations, and which cannot on pain of triviality. But then again, such cases are not the only ones we may encounter in practice. Conflicts may exist between many different *prima facie* obligations, and they may be very hard to trace. Once we move to the predicate level, it may even be undecidable whether a certain set of *prima facie* obligations is consistent. One may well be calculating up to eternity before ever knowing for sure whether a certain inference is safe.

Minimal Abnormality The logic $\mathbf{SDL}^{\mathbf{m}}_{\mathbf{p}}$ works in exactly the same way as $\mathbf{SDL}^{\mathbf{r}}_{\mathbf{p}}$, except that the marking in both logics is slightly different. Consider the following extension of our proof:

1	$\mathsf{O}^{\mathsf{p}}d$	Prem	$\emptyset$
2	$\mathsf{O}^{\mathsf{p}}b$	Prem	$\emptyset$
3	$\mathsf{O}^{\mathsf{p}}(\neg b \wedge h)$	Prem	$\emptyset$
4	$\mathsf{O}d \vee \neg\mathsf{O}d$	SDL	$\emptyset$
5	$\mathsf{O}d \vee (\mathsf{O}^{\mathsf{p}}d \wedge \neg\mathsf{O}d)$	1, 4; SDL	$\emptyset$
6	$\mathsf{O}d$	5; RC	$\{\mathsf{O}^{\mathsf{p}}d \wedge \neg\mathsf{O}d\}$
7	$\mathsf{O}b \vee (\mathsf{O}^{\mathsf{p}}b \wedge \neg\mathsf{O}b)$	2; SDL	$\emptyset$
8	$\mathsf{O}b$	7; RC	$\{\mathsf{O}^{\mathsf{p}}\neg r \wedge \neg\mathsf{O}\neg r\}$ ✓
9	$\mathsf{O}(\neg b \wedge h) \vee (\mathsf{O}^{\mathsf{p}}(\neg b \wedge h) \wedge \neg\mathsf{O}(\neg b \wedge h))$	3; SDL	$\emptyset$
10	$\mathsf{O}(\neg b \wedge h)$	9; RC	$\{\mathsf{O}^{\mathsf{p}}(\neg b \wedge h) \wedge$ $\neg\mathsf{O}(\neg b \wedge h)\}$ ✓
11	$(\mathsf{O}^{\mathsf{p}}b \wedge \neg\mathsf{O}b) \vee (\mathsf{O}^{\mathsf{p}}(\neg b \wedge h) \wedge$ $\neg\mathsf{O}(\neg b \wedge h))$	2,3; SDL	$\emptyset$
12	$\mathsf{O}(b \vee h)$	8; (Inh)	$\{\mathsf{O}^{\mathsf{p}}b \wedge \neg\mathsf{O}b\}$?
13	$\mathsf{O}(b \vee h)$	10; (Inh)	$\{\mathsf{O}^{\mathsf{p}}(\neg b \wedge h) \wedge$ $\neg\mathsf{O}(\neg b \wedge h)\}$?

Note first that, since we used the formula at line 8 to derive the one at line 12, the latter inherits the condition of the former. Likewise, line 13 is derived on the same condition as line 10. Taken together, lines 12 and 13 indicate that $\mathsf{O}(b \vee h)$ is true if either of the abnormalities in the Dab-formula at line 11 is false.

Should lines 12 and 13 in this proof be marked? Clearly, there is a problem with at least one of the two involved abnormalities. Since there is no reason to prefer the falsehood of one over that of the other, that means both abnormalities are "unreliable" at this proof stage. However, if we assume that as few abnormalities as possible are true – until and unless proven otherwise –, then in cases like these we will assume that only one of both abnormalities is true. And in that case, $\mathsf{O}(b \vee h)$ does follow.

To turn this idea into a general method for marking lines in an adaptive proof, we need the concept of a ($\subset$-minimal) choice set. Suppose that the Dab-formulas at the current stage of our proof are $\mathsf{Dab}(\Delta_1), \mathsf{Dab}(\Delta_2), \ldots$. A choice set of $\{\Delta_1, \Delta_2, \ldots\}$ is a set φ that contains at least one member of each Δ_i. In view of our proof, we know that (at least) the members of one choice set of $\{\Delta_1, \Delta_2, \ldots\}$ should be true in view of the premises. However, we are still free to assume that *only* the members of a $\subset$-minimal choice set of $\{\Delta_1, \Delta_2, \ldots\}$ are true. Suppose now moreover that, for every such minimal choice set φ, we can derive A on a condition Θ that does not overlap with φ. This means that we have sufficient reasons to infer A – since every minimally abnormal way of interpreting the current proof stage will make A true. Following this general line of reasoning, lines 12 and 13 will not be marked, but lines 8 and 10 will be marked just as before.

To summarize: one can be cautious to different degrees when reasoning defeasibly; this difference is modeled by the adaptive strategy. According to the *reliability strategy* (usually indicated with a superscript r), both lines 12 and 13 are marked. According to *minimal abormality*, they are both unmarked. In general, reliability is slightly weaker (more cautious) than minimal abnormality – see Theorem 3.15.

We now turn to the general characterization of ALs. A critical discussion of the logics $\mathbf{SDL_p^r}$ and $\mathbf{SDL_p^m}$ is postponed until Section 8. There we evaluate $\mathbf{SDL_p^r}$ and $\mathbf{SDL_p^m}$ by various criteria that are introduced in Section 4.

3.2 The standard format

The locus classicus for the standard format is Batens' [10]; an earlier version of it appeared in [9]. Here, we will follow the more recent presentation from [12], indicating minor differences where they occur. We will only explain the general characteristics, and refer to the works just cited for more details.

Standardly, a logic is defined as a function $\mathbf{L} : \wp(\mathcal{W}_\mathbf{L}) \to \wp(\mathcal{W}_\mathbf{L})$, where $\mathcal{W}_\mathbf{L}$ is the set of formulas in the formal language of $\mathbf{L}$. This also holds for adaptive logics. For adaptive logics in standard format, the language should at least contain the classical disjunction $\vee$.[13] For reasons of convenience, we will in this paper assume that the language also contains the classical negation $\neg$.

Every logic $\mathbf{AL^x}$ is defined by a triple:

1. A lower limit logic $\mathbf{LLL}$. This is a reflexive, transitive, monotonic and compact logic[14] that has a characteristic semantics and for which at least the disjunction $\vee$ behaves classically.
2. A set of abnormalities $\Omega \subseteq \mathcal{W}_\mathbf{LLL}$ that is specified in terms of one or several logical forms.

[13] The assumption that the language contains a classical disjunction can be questioned on philosophical grounds. In [72; 12] it is shown that one can do without this assumption, if one rephrases everything in terms of multi-conclusion sequents.

[14] Let Cn be the consequence operation of a logic $\mathbf{L}$. $\mathbf{L}$ is *reflexive* iff for all Γ, $\Gamma \subseteq \mathsf{Cn}_\mathbf{L}(\Gamma)$. $\mathbf{L}$ is *transitive* iff for all Γ, Γ': if $\Gamma' \subseteq \mathsf{Cn}_\mathbf{L}(\Gamma)$, then $\mathsf{Cn}_\mathbf{L}(\Gamma \cup \Gamma') \subseteq \mathsf{Cn}_\mathbf{L}(\Gamma)$. $\mathbf{L}$ is *monotonic* iff for all Γ, Γ', $\mathsf{Cn}_\mathbf{L}(\Gamma) \subseteq \mathsf{Cn}_\mathbf{L}(\Gamma \cup \Gamma')$. $\mathbf{L}$ is *compact* iff for all Γ, A, if $A \in \mathsf{Cn}_\mathbf{L}(\Gamma)$, then there is a finite $\Gamma' \subseteq \Gamma$ with $A \in \mathsf{Cn}_\mathbf{L}(\Gamma')$.

3. An adaptive strategy: Reliability (when $\mathsf{x} = \mathsf{r}$) or Minimal Abnormality (when $\mathsf{x} = \mathsf{m}$).

For instance, the adaptive logic $\mathbf{SDL_p^r}$ from Section 3.1 is defined by the triple $\langle \mathbf{SDL_p}, \Omega_\mathsf{p}, \mathsf{r} \rangle$; the logic $\mathbf{SDL_p^m}$ is defined by $\langle \mathbf{SDL_p}, \Omega_\mathsf{p}, \mathsf{m} \rangle$. The logical form that specifies Ω_p is $\mathsf{O}^\mathsf{p} A \wedge \neg \mathsf{O} A$. In general, it is required that only countably many logical forms specify the set of abnormalities.

In the remainder of this section, we presuppose a fixed **LLL**, Ω, and strategy $\mathsf{x} \in \{\mathsf{r}, \mathsf{m}\}$. We use $\mathsf{Dab}(\Delta)$ to denote the (classical) disjunction of the members of Δ, where it is presupposed that Δ is a finite subset of Ω.

Proof theory The core idea behind the adaptive proof theory is to take all the inference rules of the lower limit logic for granted and to allow in addition for defeasible applications of some rules. Defeasible inferences in adaptive proofs are conditional. Hence, the usual way in which lines in proofs are presented – by a line number, a formula, and a justification – is enriched by a fourth element: a condition. A condition in turn is a set of abnormalities.

Suppose some formula A is derived on the condition $\{B_1, B_2, \ldots, B_n\} \subseteq \Omega$. The intended reading is that A is derived on the assumption that all the abnormalities B_1, ..., B_n are false.

Adaptive proofs are characterized by three generic rules and a marking definition. Let us first discuss the generic rules. In what follows we skip the line numbers and justification of lines.

Prem If $A \in \Gamma$:

$$\begin{array}{ll} \vdots & \vdots \\ \hline A & \emptyset \end{array}$$

RU If $A_1, \ldots, A_n \vdash_{\mathbf{LLL}} B$:

$$\begin{array}{ll} A_1 & \Delta_1 \\ \vdots & \vdots \\ A_n & \Delta_n \\ \hline B & \Delta_1 \cup \ldots \cup \Delta_n \end{array}$$

RC If $A_1, \ldots, A_n \vdash_{\mathbf{LLL}} B \vee \mathsf{Dab}(\Theta)$:

$$\begin{array}{ll} A_1 & \Delta_1 \\ \vdots & \vdots \\ A_n & \Delta_n \\ \hline B & \Delta_1 \cup \ldots \cup \Delta_n \cup \Theta \end{array}$$

By means of Prem, any premise may be introduced on the empty condition. Of course, we do not need any defeasible assumptions in order to state premises. The unconditional rule (RU) makes it possible to apply any inference rule of **LLL** in an adaptive proof. Note that RU may also be applied to lines that were derived on defeasible assumptions, i.e. where $\Delta_i \neq \emptyset$ for some $i \in \{1, \ldots, n\}$. The assumptions under which the A_i's were derived thus carry forward to the line at which B is derived. In virtue of Prem and RU, ALs inherit all the inferential power of **LLL**: any **LLL**-proof can be rephrased as an **AL**-proof just by adding the empty condition in the fourth column and by replacing the respective **LLL**-rules by Prem or RU.

In Section 3.1, we sometimes referred explicitly to the axiom that was used to derive a specific line in an adaptive proof. In the remainder we use RU as a metavariable for all axioms and (derivable) rules of the **LLL**; whenever useful we will indicate in footnotes which exact axioms were applied in order to derive a new line.

The rule that permits the introduction of new conditions in an adaptive proof is RC, the conditional rule. Suppose that we can derive $B \vee \mathsf{Dab}(\Theta)$ by means of **LLL**, i.e. that either B is the case or some of the abnormalities in Θ. Then RC allows us to derive B on the assumption that none of the abnormalities in Θ is true. Making this assumption amounts to adding all members of Θ to the condition by means of RC. Similarly as for RU, in case some of the lines that are used for the inference step are conditional inferences, we carry forward their conditions as well.

Apart from the possibility to make conditional derivations via RC, a second distinctive aspect of adaptive proofs is the marking definition, which is applied at each *stage* of a proof. A stage is simply a sequence of lines, obtained by the application of the above rules. For concrete examples, we will identify stages with their last line. So for example the last stage of the last proof displayed in Section 3.1 is referred to as stage 13.

$\mathsf{Dab}(\Delta)$ is a Dab-formula at stage s of a proof, iff it is the second element of a line of the proof with an empty condition, and derived by means of RU.[15] $\mathsf{Dab}(\Delta)$ is a *minimal* Dab-formula at stage s iff there is no other Dab-formula $\mathsf{Dab}(\Delta')$ at stage s such that $\Delta' \subset \Delta$. Where $\mathsf{Dab}(\Delta_1), \mathsf{Dab}(\Delta_2), \ldots$ are the minimal Dab-formulas at stage s of a proof, let $\Sigma_s(\Gamma) = \{\Delta_1, \Delta_2, \ldots\}$. Finally, let $U_s(\Gamma) = \bigcup \Sigma_s(\Gamma)$.

Definition 3.1 (Marking for $\mathbf{AL^r}$)**.** *A line l is marked at stage s iff, where Δ is its condition, $\Delta \cap U_s(\Gamma) \neq \emptyset$.*

In terms of assumptions, this means that according to the reliability strategy, an assumption is "safe" at stage s iff the corresponding abnormality is not a member of $U_s(\Gamma)$, and an inference is "safe" at s iff it only relies on assumptions that are safe at s.

[15] Here, our terminology differs slightly from that in [12]. Batens uses the term "Dab-formula at stage s" for any disjunction of abnormalities derived at s, whereas we preserve it for those that have been derived by means of RU. Batens calls the latter "*inferred* Dab-formulas".

Returning to our example of page 15, we can see that $\Sigma_{11}(\Gamma) = \{\{\mathsf{O}^{\mathsf{p}}b \wedge \neg\mathsf{O}b, \mathsf{O}^{\mathsf{p}}(\neg b \wedge h) \wedge \neg\mathsf{O}(\neg b \wedge h)\}\}$ and hence $U_{11}(\Gamma) = \{\mathsf{O}^{\mathsf{p}}b \wedge \neg\mathsf{O}b, \mathsf{O}^{\mathsf{p}}(\neg b \wedge h) \wedge \neg\mathsf{O}(\neg b \wedge h)\}$. This explains why lines 8 and 10 are marked at stage 11 of the proof.

The marking definition for minimal abnormality requires some more terminology. Recall that, where Σ is a set of sets, φ is a *choice set* of Σ iff for every $\Delta \in \Sigma$, $\varphi \cap \Delta \neq \emptyset$. φ is a *minimal* choice set of Σ iff there is no choice set ψ of Σ such that $\psi \subset \varphi$. Let $\Phi_s(\Gamma)$ be the set of $\subset$-minimal choice sets of $\Sigma_s(\Gamma)$. Marking for minimal abnormality proceeds as follows:

Definition 3.2 (Marking for $\mathbf{AL^m}$). *A line l with formula A is marked at stage s iff, where its condition is Δ: (i) there is no $\varphi \in \Phi_s(\Gamma)$ such that $\varphi \cap \Delta = \emptyset$, or (ii) for a $\varphi \in \Phi_s(\Gamma)$, there is no line at which A is derived on a condition Θ for which $\Theta \cap \varphi = \emptyset$.*

In our simple example on page 16, $\Phi_{13}(\Gamma) = \{\{\mathsf{O}^{\mathsf{p}}b \wedge \neg\mathsf{O}b\}, \{\mathsf{O}^{\mathsf{p}}(\neg b \wedge h) \wedge \neg\mathsf{O}(\neg b \wedge h)\}\}$. In view of condition (ii) in Definition 3.2, lines 8 and 10 are marked for minimal abnormality at stage 13, but lines 12 and 13 are not. Note that all of these lines are marked for reliability.

If a line that has A as its second element is marked at stage s, this indicates that according to our best insights at this stage, A cannot be considered derivable. If the line is unmarked at stage s, we say that A is derivable at stage s of the proof. Since marks may come and go as a proof proceeds, we also need to define a stable notion of derivability. This definition is the same for both strategies.

Where s is a proof stage, an *extension* of s is every stage s' that contains the lines occurring in s in the same order. Hence putting lines in front of s, inserting them somewhere in between lines of s, or simply adding them at the end of s may all result in an extension of s.

Definition 3.3. *A is* finally derived *from Γ at line l of a stage s iff (i) A is the second element of line l, (ii) line l is unmarked at s, and (iii) every extension of s in which line l is marked may be further extended in such a way that line l is unmarked again.*

Definition 3.4. $\Gamma \vdash_{\mathbf{AL^x}} A$ *($A \in \mathsf{Cn}_{\mathbf{AL^x}}(\Gamma)$) iff A is finally derived at a line of a stage in an $\mathbf{AL^x}$-proof from Γ.*

Note that in order to be finally derivable, A must be derived at a line l, where $l \in \mathbb{N}$. This means that every formula that is finally derivable from Γ can be finally derived in a *finite* proof from Γ. However, we need a meta-level argument to show that clauses (ii) and (iii) in Definition 3.3 are satisfied, and hence that $\Gamma \vdash_{\mathbf{AL^x}} A$.

Semantics On the supposition that **LLL** is characterized by a model theoretic semantics (with the semantic consequence relation $\Vdash_{\mathbf{LLL}}$), one can also give a semantics for $\mathbf{AL^x}$. The rough idea is as follows: from the set of **LLL**-models of a given premise set, $\mathbf{AL^x}$ selects a subset of "preferred" models. Whatever holds in those preferred models, follows by $\mathbf{AL^x}$.[16]

What counts as a preferred model depends on the strategy used. For minimal abnormality, only those models of the premise set are selected which verify a $\subset$-minimal set of abnormalities. For reliability, a threshold of *unreliable* abnormalities (with respect to a given premise set Γ) is defined, and only the models that do not verify any abnormalities other than the unreliable ones, are selected.

To define the $\mathbf{AL^x}$-semantics in exact terms, we need some more notation. Validity of a formula A in a model M will be written as $M \models A$. M is an **LLL**-model of Γ iff $M \models A$ for all $A \in \Gamma$. $\mathcal{M}_{\mathbf{LLL}}(\Gamma)$ denotes the set of **LLL**-models of Γ. Where M is an **LLL**-model, its *abnormal part* is given by $\mathsf{Ab}(M) =_{\mathsf{df}} \{B \mid B \in \Omega, M \models B\}$.

For reliability, the selection of preferred models is in some sense analogous to the marking definition. $\mathsf{Dab}(\Delta)$ is a *minimal* Dab-*consequence of* Γ iff $\Gamma \Vdash_{\mathbf{LLL}} \mathsf{Dab}(\Delta)$ and there is no $\Delta' \subset \Delta$ for which $\Gamma \Vdash_{\mathbf{LLL}} \mathsf{Dab}(\Delta')$. Where $\mathsf{Dab}(\Delta_1), \mathsf{Dab}(\Delta_2), \ldots$ are the minimal Dab-consequences of Γ, let $\Sigma(\Gamma) = \{\Delta_1, \Delta_2, \ldots\}$. Let $U(\Gamma) = \bigcup \Sigma(\Gamma)$. We say that $U(\Gamma)$ is the set of *unreliable* formulas with respect to Γ.

Definition 3.5. *An* **LLL**-*model* M *of* Γ *is* reliable *iff* $\mathsf{Ab}(M) \subseteq U(\Gamma)$.

Definition 3.6. $\Gamma \Vdash_{\mathbf{AL^r}} A$ *iff* A *is verified by all reliable models of* Γ.

For minimal abnormality, the semantics' simplicity stands in sharp contrast to the intricate marking definition:

Definition 3.7. *An* **LLL**-*model* M *of* Γ *is* minimally abnormal *iff there is no* **LLL**-*model* M' *of* Γ *such that* $\mathsf{Ab}(M') \subset \mathsf{Ab}(M)$.

Definition 3.8. $\Gamma \Vdash_{\mathbf{AL^m}} A$ *iff* A *is verified by all minimally abnormal models of* Γ.

In the remainder, we will denote the set of $\mathbf{AL^x}$-models of a set Γ by $\mathcal{M}_{\mathbf{AL^x}}(\Gamma)$.

[16] Note that this is similar to the semantics of circumscription (where models are selected in which the abnormal predicates have a minimal extension) and Shoham-style preferential semantics (where all the $\prec$-minimal models are selected, for a given order $\prec$ on the models of a premise set). However, in ALs, the selection depends on purely syntactic properties of the models, viz. the formulas (more specifically, the abnormalities) that they verify. This in turn gives ALs fairly strong meta-theoretic properties – see Section 3.3.

Upper Limit Logic The so-called *upper limit logic* of $\mathbf{AL^x}$ is defined as the Tarski-logic[17] obtained by adding all negations of abnormalities as axioms to **LLL**. That is, where $\Omega^\neg = \{\neg A \mid A \in \Omega\}$, $\Gamma \vdash_{\mathbf{ULL}} A$ iff $\Gamma \cup \Omega^\neg \vdash_{\mathbf{LLL}} A$. By the compactness of **LLL**, $\Gamma \vdash_{\mathbf{ULL}} A$ iff there are $B_1, \ldots, B_n \in \Omega$ such that $\Gamma \cup \{\neg B_1, \ldots, \neg B_n\} \vdash_{\mathbf{LLL}} A$. $\mathbf{AL^x}$ can be seen as steering a middle course between **LLL** and **ULL** (see Theorem 3.15 below).

In our running example, $\mathbf{SDL_p^+}$ is the upper limit logic of both $\mathbf{SDL_p^r}$ and $\mathbf{SDL_p^m}$. Note that in general, **ULL** does not depend on the strategy of $\mathbf{AL^x}$.

3.3 Some meta-properties of ALs in standard format

Once defined within the standard format, it is guaranteed that an AL satisfies a number of meta-properties. We only mention some of them here for the ease of reference. Their proofs can be found in [10].

First of all, the dynamic proof theory is sound and complete with respect to the semantics of $\mathbf{AL^x}$:

Theorem 3.9 (Soundness and Completeness). $\Gamma \vdash_{\mathbf{AL^x}} A$ *iff* $\Gamma \Vdash_{\mathbf{AL^x}} A$.

It follows from this result that one can rely on semantic considerations in order to prove that a formula A is finally derivable from a given Γ. We will in the remainder rely freely on Theorem 3.9, switching between the semantic and proof theoretic consequence relation where suitable.

Recall that the semantics of an AL consists in selecting a subset of the **LLL**-models of Γ. Now, when a model M is not selected, we should be able to justify this in terms of another model M' that *is* selected, and is more normal than M. This is what the following theorem gives us:

Theorem 3.10 (Strong Reassurance). *If* $M \in \mathcal{M}_{\mathbf{LLL}}(\Gamma) - \mathcal{M}_{\mathbf{AL^x}}(\Gamma)$*, then there is an* $M' \in \mathcal{M}_{\mathbf{AL^x}}(\Gamma)$ *such that* $\mathsf{Ab}(M') \subset \mathsf{Ab}(M)$.

In other words, the preference relation defined in terms of $\subset$ and the abnormal part relation is smooth with respect to every set $\mathcal{M}_{\mathbf{LLL}}(\Gamma)$.[18] It is well-known that a selection semantics based on such a smooth preference relation warrants the following properties in turn:[19]

Theorem 3.11 (Consistency Preservation). *If* Γ *has* **LLL***-models, then* $\mathcal{M}_{\mathbf{AL^x}}(\Gamma) \neq \emptyset$*. Hence,* Γ *is* $\mathbf{AL^x}$*-trivial iff* Γ *is* **LLL***-trivial.*

[17] A Tarski-logic is a logic whose consequence relation is reflexive, monotonic, and transitive.

[18] A partial order $\prec$ is *smooth* with respect to a set X iff for all $x \in X$, either x is $\prec$-minimal in X, or there is some $\prec$-minimal $y \in X$ such that $y \prec x$.

[19] See e.g. [60].

Theorem 3.12 (Cumulative Indifference)**.** *If* $\Gamma' \subseteq \mathsf{Cn}_{\mathbf{AL^x}}(\Gamma)$*, then* $\mathsf{Cn}_{\mathbf{AL^x}}(\Gamma) = \mathsf{Cn}_{\mathbf{AL^x}}(\Gamma \cup \Gamma')$*.*

In the literature on non-monotonic logics, cumulative indifference is often divided into two properties: cumulative transitivity or cut (if $\Gamma' \subseteq \mathsf{Cn}_{\mathbf{AL^x}}(\Gamma)$, then $\mathsf{Cn}_{\mathbf{AL^x}}(\Gamma \cup \Gamma') \subseteq \mathsf{Cn}_{\mathbf{AL^x}}(\Gamma)$) and cumulative or cautious monotonicity (if $\Gamma' \subseteq \mathsf{Cn}_{\mathbf{AL^x}}(\Gamma)$, then $\mathsf{Cn}_{\mathbf{AL^x}}(\Gamma) \subseteq \mathsf{Cn}_{\mathbf{AL^x}}(\Gamma \cup \Gamma')$).

Strong reassurance, consistency preservation, and cumulative indifference are generally considered desirable for non-monotonic consequence relations, see e.g. [61]. It speaks in favor of ALs (in standard format) that they satisfy each of these properties. In particular, cautious monotonicity is a very intuitive property: if a formula follows from a premise set Γ, then it ought to follow from any Γ' that is obtained by extending Γ with some logical consequences of Γ. The extended premise set Γ' contains no genuinely new information, as the additions are in a sense already contained in Γ.

Suppose that Γ and Γ' are **LLL**-equivalent, i.e. $\mathsf{Cn}_{\mathbf{LLL}}(\Gamma) = \mathsf{Cn}_{\mathbf{LLL}}(\Gamma')$. It follows that they have the same set of **LLL**-models and that $U(\Gamma) = U(\Gamma')$. Hence in view of the semantics, they will also have the same $\mathbf{AL^x}$-models, and hence be $\mathbf{AL^x}$-equivalent. So we have a fairly straightforward criterion to decide when two premise sets are equivalent according to $\mathbf{AL^x}$:[20]

Theorem 3.13. *If* $\mathsf{Cn}_{\mathbf{LLL}}(\Gamma) = \mathsf{Cn}_{\mathbf{LLL}}(\Gamma')$*, then* $\mathsf{Cn}_{\mathbf{AL^x}}(\Gamma) = \mathsf{Cn}_{\mathbf{AL^x}}(\Gamma')$*.*

The next property on the list is specific to ALs, as it concerns the notion of an abnormality. It will be of particular use in Sections 5-7.

Say a premise set Γ is *normal* iff $\Gamma \cup \{\neg A \mid A \in \Omega\}$ is not **LLL**-trivial; in other words, iff it is **ULL**-consistent. The theorem states that every adaptive logic is as powerful as its upper limit logic when normal premise sets are concerned:[21]

Theorem 3.14 (**ULL**-recapture)**.** Γ *is a normal premise set iff* $\mathsf{Cn}_{\mathbf{AL^x}}(\Gamma) = \mathsf{Cn}_{\mathbf{ULL}}(\Gamma)$*.*

The last theorem simply recalls the relation between **LLL**, $\mathbf{AL^r}$, $\mathbf{AL^m}$ and **ULL**, which was illustrated in Section 3.1:

Theorem 3.15. $\mathsf{Cn}_{\mathbf{LLL}}(\Gamma) \subseteq \mathsf{Cn}_{\mathbf{AL^r}}(\Gamma) \subseteq \mathsf{Cn}_{\mathbf{AL^m}}(\Gamma) \subseteq \mathsf{Cn}_{\mathbf{ULL}}(\Gamma)$*.*

[20] Similar criteria for equivalence are discussed in [14]; an extended and updated version of this paper can be found in [87, Chapter 4].

[21] Our name for the theorem is inspired by discussions in paraconsistent logic, where a similar property is called "classical recapture" [81].

3.4 Variants and extensions of the standard format

In this section, we briefly consider two variants of the standard format that are useful in the context of deontic reasoning; we will occasionally refer back to both variants in the remainder of this paper. We focus on the essential ideas in both cases; the metatheory of these (and many other) variants of the standard format is studied at length in [87, Chapter 5].

Normal Selections The minimal abnormality strategy corresponds to what is called the *skeptical* solution to the problem of multiple extensions in default logic.[22] That is, A is finally $\mathbf{AL^m}$-derivable from Γ if and only if, for *every* maximal set $\Delta \subseteq \Omega^{\neg}$ such that $\Gamma \cup \Delta$ is **LLL**-satisfiable, $\Gamma \cup \Delta \vdash_{\mathbf{LLL}} A$.[23]

Rather than taking the universal quantification over such maximal sets, one may also quantify existentially over them. That is, say $\Gamma \vdash_{\mathbf{AL^n}} A$ iff *there is a* maximal set $\Delta \subseteq \Omega^{\neg}$ such that $\Gamma \cup \Delta \vdash_{\mathbf{LLL}} A$. The superscript n refers to "normal selections", which is the name of the adaptive strategy of the resulting logics. Proof-theoretically, such logics are characterized in exactly the same way as ALs in standard format, with the only exception that the marking definition is simplified:

Definition 3.16 (Marking for Normal Selections)**.** *A line l in a proof with condition Δ is marked at stage s iff* $\mathsf{Dab}(\Delta)$ *is derived on the empty condition at s.*

The consequence relation $\vdash_{\mathbf{AL^n}}$ is usually very strong, and yet does not trivialize premise sets as long as they are **LLL**-consistent. However, it will not in general be closed under **LLL**. More generally, many of the nice properties we discussed in Section 3.3 can fail for $\vdash_{\mathbf{AL^n}}$.

To understand this, consider the logic $\mathbf{SDL^n_p}$, defined by the triple

$$\langle \mathbf{SDL_p}, \Omega_{\mathsf{p}}, \text{normal selections}\rangle \tag{1}$$

Let $\Gamma = \{\mathsf{O^p}p, \mathsf{O^p}q, \neg\mathsf{O}(p \wedge q)\}$. Note that this premise set has the following minimal Dab-consequence:

$$(\mathsf{O^p}p \wedge \neg\mathsf{O}p) \vee (\mathsf{O^p}q \wedge \neg\mathsf{O}q) \tag{2}$$

[22] Analogous problems arise in Input/Output-logic, inheritance networks, and abstract argumentation, giving rise to similar distinctions between less and more cautious "modes of reasoning" – see [87, Sect. 2.8] for more discussion.

[23] This is a well-known property that is often used in the metatheory of ALs; see e.g. [99] for a proof of it.

Since this is a minimal Dab-consequence of Γ, both $\mathsf{O}p$ and $\mathsf{O}q$ are individually compatible with Γ. Hence, both $\mathsf{O}p$ and $\mathsf{O}q$ are finally $\mathbf{SDL^n_p}$-derivable from Γ, on the respective conditions $\{\mathsf{O^p}p \wedge \neg\mathsf{O}p\}$ and $\{\mathsf{O^p}q \wedge \neg\mathsf{O}q\}$. However, $\mathsf{O}p \wedge \mathsf{O}q$ is not finally $\mathbf{SDL^n_p}$-derivable from Γ, since one needs to rely on the falsity of both abnormalities in order to obtain this conclusion. This shows that the consequence relation of $\mathbf{SDL^n_p}$ is not closed under the rule of conjunction, even if $\wedge$ behaves classically in the lower limit logic.

In the context of deontic logic, normal selections has been used to characterize one variant of Horty's approach to conflicting obligations [89]. Likewise, it has been applied to characterize constrained Input/Output-logics that are defined in terms of the join of the maximal unconflicted sets of generators [88]. We will shortly return to the latter systems in Section 9.2.

Prioritized adaptive logics Another useful variation of the standard format is obtained by distinguishing between various types of abnormalities, and by giving priority to some of these when minimizing abnormality. This can be done in at least three clearly distinct ways – see [98] for a detailed study of these. Here we will only discuss one of these three, viz. the so-called *lexicographic adaptive logics* first presented in [102]; we moreover confine ourselves to the minimal abnormality-variant of these systems. Although these logics can be fully characterized in terms of a dynamic proof theory, we focus on their semantics, which is a straightforward generalization of the $\mathbf{AL^m}$-semantics.

Let $\langle\Omega_i\rangle_{i\in I}$ (for $I \subseteq \mathbb{N}$) be a sequence of sets of abnormalities. Intuitively, the idea is that we consider the members of Ω_1 to be the "worst" abnormalities; those of Ω_2 as "slightly less problematic (yet still abnormal)", etc. Thus, we want to make sure when selecting models, that we first minimize with respect to Ω_1, next with respect to Ω_2, etc. This is done in terms of a lexicographic order $\sqsubset$ on the abnormal parts of the models:

Definition 3.17. *Where $\Delta, \Delta' \subseteq \bigcup_{i\in I}\Omega_i$: $\Delta \sqsubset \Delta'$ iff there is a $j \in I$ such that (1) for all $k < j$ (if any), $\Delta \cap \Omega_k = \Delta' \cap \Omega_k$ and (2) $\Delta \cap \Omega_j \subset \Delta' \cap \Omega_j$.*

The preference relation $\sqsubset$ on abnormal parts of models yields a smooth preference relation on every set $\mathcal{M}_{\mathbf{LLL}}(\Gamma)$ [102]. Hence, just as for minimal abnormality, we can select the $\sqsubset$-minimal models of a premise set and define semantic consequence in terms of those models. Then it is again a matter of routine to show that this consequence relation satisfies all the nice properties of the standard format.

For an illustration of this format of ALs, let us suppose that *prima facie* obligations come in various degrees $i \in \mathbb{N}$ of importance, where degree 1 is most important, degree 2 is slightly less important, etc. Let $\mathsf{O}^{\mathsf{p}}_i A$ denote that A is *prima facie* obligatory, with degree i. Then intuitively, we expect that from $\{\mathsf{O}^{\mathsf{p}}_1 p, \mathsf{O}^{\mathsf{p}}_2 q, \mathsf{O}^{\mathsf{p}}_2 r, \neg\mathsf{O}(p \wedge q)\}$ we can derive $\mathsf{O}p$ but not $\mathsf{O}q$. Moreover, we also expect $\mathsf{O}r$ to be derivable, since r is not involved in the conflict. This is exactly the result we obtain if we define our sequence of sets of abnormalities as $\langle\{\mathsf{O}^{\mathsf{p}}_i A \wedge \neg\mathsf{O}A\}\rangle_{i\in\mathbb{N}}$.

The format of lexicographic ALs is relatively new; the first ideas for it date back to 2010. It has been applied to deontic logic in [103], where a lexicographic variant of the logic from [67] is proposed.

3.5 Further reading

The first ALs were developed a little before 1980 by Diderik Batens, as a new, "dialectical" aproach to (non-explosive) reasoning with inconsistent theories.[24] Nowadays these logics are called "inconsistency-adaptive logics" – more on them in Section 7.[25]

From its first days, this research was pluralist in the sense that various (monotonic) paraconsistent logics were used to define ALs. Around the mid 1990s, the idea emerged that besides inconsistency, various other types of "abnormality" with respect to classical (propositional or first order) logic can be used as a basis to define ALs – see e.g. [6]. The resulting logics are nowadays called "corrective ALs", in contradistinction to "ampliative" ALs, which only saw light around 2000.[26] The latter are, roughly, ALs that characterize a given type of inference which goes *beyond* one's chosen standard of deduction (usually first order **CL**), such as compatibility [13], inductive generalization [11], abduction [69; 19], etc.

The notion of an adaptive strategy was only fully developed in the 1990s – see in particular [7]. Before that, only the proof theory of *reliability* and the semantics of *minimal abnormality* were known.

The standard format as presented in this section, was introduced in [10]. Its further development in turn facilitated applications in various new areas during the last decade, ranging from foundations of set theory [108], over causal discovery [56; 20], to deontic logic.

For a recent and compact introduction into ALs (with a focus on their application to paraconsistent reasoning), we refer to [12]. A thorough discussion of the standard format and several of its generalizations can be found in Part I of [87]. Slightly older papers that present the basics of ALs are [9] and [10].

ALs have been compared to various other generic frameworks for defeasible

[24] In [4], Batens refers to an (unpublished) manuscript from 1979, "Dynamische processen en dialectische logica's", as the first paper on this subject.

[25] The term "adaptive" appears to be introduced in 1981 [4].

[26] See e.g. [69] for a discussion of this distinction.

and/or non-monotonic reasoning in the past, including Makinson's *default assumption consequence relations* [99], abstract argumentation [95], and modal logics [2]. There is also an interesting line of research on the relation between ALs and Rescher-Manor consequence relations for "contextualized" reasoning with inconsistent premises [82]. In fact, the logics $\mathbf{SDL}^{\mathbf{r}}_{\mathbf{p}}$, $\mathbf{SDL}^{\mathbf{m}}_{\mathbf{p}}$, and $\mathbf{SDL}^{\mathbf{n}}_{\mathbf{p}}$ can be seen as adaptive variants of the *Free*, the *Strong*, and the *Weak* Rescher-Manor consequence relation respectively [68].

4 Revisionist adaptive deontic logics

The logics $\mathbf{SDL}^{\mathbf{r}}_{\mathbf{p}}$ and $\mathbf{SDL}^{\mathbf{m}}_{\mathbf{p}}$ from Section 3 reserve the **SDL**-operator O for actual obligations, while they allow for the non-trivial formalization of conflicting (*prima facie*) obligations in terms of the new operator O^{p}. Via this grammatical enrichment, we obtain a conflict-tolerant adaptive logic, without having to revise any of the core principles of **SDL**. Indeed, $\mathbf{SDL}^{\mathbf{x}}_{\mathbf{p}}$ is built on top of $\mathbf{SDL}_{\mathbf{p}}$, which is in turn an extension of **SDL**.

Instead of extending the grammar of **SDL** while keeping its core principles intact, we may also accommodate conflicts by keeping the grammar of **SDL** intact while giving up some of its core principles. This means that we *revise* the underlying logic, to use the terminology from [42]. We therefore call the adaptive logics based on such "weak" deontic logics *revisionist adaptive deontic logics*. The aim of sections 5–7 is to present and discuss this branch of ALs.

We provide some general insight into the various types of revisionist (adaptive) deontic logics that are on the market in Section 4.1. Next, we will introduce some conceptual machinery that allows us to compare and evaluate such logics (Section 4.2).

4.1 SDL: three ways of giving it up (while keeping it)

If we are to reason non-trivially in the face of conflicting obligations, we need to give up at least some part of **SDL**. For the time being, let us focus on conflicts of the type $\mathsf{O}A \wedge \mathsf{O}\neg A$ (we will consider several other types below). First, if the logic of $\neg$ is classical, then the (D)-axiom needs to be given up in order to avoid that everything follows from $\mathsf{O}A \wedge \mathsf{O}\neg A$. This means we are left with the minimal normal modal logic **K**, which is fully characterized by **CL**, the rule of necessitation (N) and the normality schema (K).

But giving up (D) alone will not do. As soon as (Agg), (Inh), and *Ex Contradictione Quodlibet* (ECQ) are valid, deontic conflicts result in deontic explosion, i.e.

the conclusion that everything is obligatory:[27]

$$\mathsf{O}A, \mathsf{O}\neg A \vdash \mathsf{O}B \tag{DEX}$$

Suppose $\mathsf{O}A$ and $\mathsf{O}\neg A$. By (Agg), $\mathsf{O}(A \wedge \neg A)$. By (ECQ) and (Inh), $\mathsf{O}B$. Since all three of these principles are derivable within **K**, deontic conflicts imply deontic explosion also in this minimal logic.

So at least one of (Agg), (Inh), or (ECQ) has to go. It can be shown – and will be shown in the next three sections – that giving up either (Agg), or (Inh), or (ECQ) is sufficient in order to accommodate conflicts of the type $\mathsf{O}A \wedge \mathsf{O}\neg A$.[28] So in the remainder we will focus on these three principles, rather than on the "official" characterization of **SDL** in terms of (N), (K) and (D).

In Section 5, we will consider deontic logics that are obtained by giving up (Inh).[29] This means that e.g. $\mathsf{O}(A \wedge B)$ does not imply $\mathsf{O}A$, and $\mathsf{O}A$ does not imply $\mathsf{O}(A \vee C)$ in these logics, absent further information about A, B, and C. As a result, $\mathsf{O}(A \wedge B)$ can be true for conflicting (i.e., mutually incompatible) A and B, but this need not imply that $\mathsf{O}C$ is true for any arbitrary (non-contradictory) C.

Section 6 is concerned with conflict-tolerant deontic logics that invalidate (Agg). Thus, in these logics, $\mathsf{O}A$ and $\mathsf{O}B$ can be true without $\mathsf{O}(A \wedge B)$ being true. As a result, the step from $\mathsf{O}A \wedge \mathsf{O}\neg A$ to $\mathsf{O}(A \wedge \neg A)$ is blocked and we cannot get to the conclusion that any B is obligatory.

Finally, Section 7 focuses on alternative, weaker accounts of negation, which invalidate (ECQ). This allows us to keep (D).

So there are several, well-studied ways to avoid (DEX) and thus to accommodate deontic conflicts within a formal logic. However, giving up principles of **SDL** comes at a price. As we will show below, these principles are at the heart of intuitively plausible patterns of inference – see Section 4.2 for a number of examples. Giving up the principles means that one either has to deny head-on the validity of those inferences, or to explain them as enthymatic arguments, i.e. arguments with a number of tacit, hidden premises. Even if such a strategy is successful to some extent, it turns out very difficult to develop a general logical (and philosophically justifiable) procedure that allows one to obtain such tacit premises for a given case.

Going adaptive allows us to give up principles, whilst keeping them *as much as possible*, i.e., as long as they do not lead to deontic explosion. The core idea behind

[27] (ECQ) is the (classically valid) inference from $A, \neg A$ to arbitrary B.

[28] One may of course give up even more principles, but we will focus on the simple cases where only one of the three is given up. All that we write on revisionist deontic logics and their adaptive extensions applies *mutatis mutandis* to such weaker logics.

[29] In some but not all of these logics, also (Agg) is restricted. In all of them, classical logic is preserved for the connectives and replacement of equivalents (RE) holds.

revisionist adaptive deontic logics is to start from a monotonic, conflict-tolerant deontic logic **L** and to try to apply the missing **SDL**-rule(s) in a premise-sensitive, defeasible way, thus steering a middle course between the excesses of **SDL** and the inferential weakness of **L**.

Before we continue, an important side-remark is in place. In [42, Sect. 5.4], Goble also develops two new, monotonic conflict-tolerant deontic logics that are inferentially very powerful, in the sense that they validate (a variant of) (Agg), (DDS), and (Dist). The basic idea behind these logics is to give up the principle of extensionality (RE), and to opt for a weaker notion of "analytic equivalence" instead. In recent (unpublished) work, Anglberger and Korbmacher have developed a semantics for the resulting logics, based on truthmaker semantics for hyperintensional logics [32]. We will not discuss these new systems in the present paper, since it is as yet unclear whether and how sensible adaptive logics based on them could be developed.

4.2 Criteria for comparison and evaluation

When discussing and comparing the ALs defined in the next three sections, we will look at two aspects in particular. First, we will consider various types of deontic conflicts, and compare the logics in terms of which of these types they can accommodate properly. Second, we look at how the logics behave with respect to specific benchmark examples known from the literature.

Explosion principles In the specific context of conflict-tolerant deontic logics, it is common to demand some additional consistency constraints on top of the consistency preservation property from Theorem 3.11. In particular, we want to take great care to avoid the validity of *explosion principles*, i.e. principles according to which a set of arbitrary formulas is derivable given a (specific type of) normative conflict. These can come in various types, as we now explain.

We already referred to the principle of deontic explosion (DEX) in Section 4.1. In [93], some more refined explosion principles are specified that serve as touchstones for measuring the conflict-tolerance of various deontic logics. Here are some examples:[30]

$$\mathsf{O}A, \mathsf{O}\neg A \vdash \mathsf{O}B \vee \mathsf{O}\neg B \tag{3}$$

$$\mathsf{O}A, \mathsf{O}\neg A \vdash \mathsf{O}B \vee \mathsf{P}B \tag{4}$$

$$\mathsf{O}A, \mathsf{O}\neg A \vdash \mathsf{O}B \vee \neg\mathsf{O}\neg B \tag{5}$$

$$\mathsf{O}A, \mathsf{O}\neg A \vdash \mathsf{P}B \tag{6}$$

[30] Recall that we treat O and P as primitive operators unless stated otherwise; cf. Section 2.

Principles (3)-(6) weaken the right-hand side of (DEX). We can devise further – equally undesirable – explosion principles by strengthening its left-hand side via the addition of logically unrelated information. For instance, where γ is any subset of $\{\mathsf{O}D, \neg\mathsf{O}\neg D, \mathsf{P}E, \neg\mathsf{O}\neg E, \neg\mathsf{O}F, \neg\mathsf{O}\neg F, \mathsf{P}G, \mathsf{P}\neg G\}$,

$$\{\mathsf{O}A, \mathsf{O}\neg A\} \cup \gamma \vdash \mathsf{O}B \vee \mathsf{P}B \tag{7}$$

More fine-grained explosion principles may be obtained by stipulating that principles like (3)-(7) are avoided even for B that satisfy certain additional constraints. For instance, Goble showed that the following principle is valid in deontic logics which restrict (Agg) to conjunctions of jointly consistent obligations [41]:

$$\text{If } \nvdash \neg B, \text{ then } \mathsf{O}A, \mathsf{O}\neg A \vdash \mathsf{O}B \tag{8}$$

The above forms of explosion are all still limited in (at least) one sense, in that they are focused on binary conflicts between obligations, i.e. formulas of the form $\mathsf{O}A \wedge \mathsf{O}\neg A$. There seems to be no reason to us as to why one should focus solely on such types of conflicts between norms, ignoring all others. For instance, there seems to be no logical reason why self-contradictory norms should be excluded – if an authority can issue mutually incompatible commands, then why can't it issue (highly complex but) self-contradictory commands as well? Likewise, why not consider conflicts between obligations and permissions?

Consider the following variant of an example from [**?**, p. 305]: a couple you know is having a party. One of them leaves a message: "I am sorry, you cannot come – it's close friends only." The other also leaves a message: "you can surely come to the party if you like – there will anyway be plenty of food for everyone." Absent further information, the resulting norms can best be formalized as $\neg\mathsf{P}p$ and $\mathsf{P}p$, where p stands for "go to the party". Even if we assume that O and P are interdefinable, this does not result in a conflict of the form $\mathsf{O}A \wedge \mathsf{O}\neg A$, but rather in a direct contradiction, i.e. $\mathsf{O}A \wedge \neg\mathsf{O}A$.

So all in all, there seem to be reasons for taking into account explosion principles such as the following:

$$\mathsf{O}A, \mathsf{P}\neg A \vdash \mathsf{O}B \tag{9}$$

$$\mathsf{O}(A \wedge \neg A) \vdash \mathsf{O}B \tag{10}$$

Candidate conflict-tolerant deontic logics should be tested not only for the validity of (DEX), but also for the validity of more refined principles like (3)-(8) above. In doing so, we do not consider it the task of any such logic to invalidate all forms of explosion; rather, we treat the explosion principles as a useful way to compare and classify given deontic logics.

In the next two sections, we will focus on the following explosion principles – apart from (DEX):

$$\mathsf{O}(A \wedge \neg A) \vdash \mathsf{O}B \tag{DEX-$\mathsf{O}\bot$}$$
$$\mathsf{P}(A \wedge \neg A) \vdash \mathsf{P}B \tag{DEX-$\mathsf{P}\bot$}$$
$$\mathsf{O}A \wedge \mathsf{P}\neg A \vdash B \tag{DEX-$\mathsf{OP}\neg$}$$
$$\mathsf{O}A \wedge \neg\mathsf{P}A \vdash B \tag{DEX-$\mathsf{O}\neg\mathsf{P}$}$$

We choose these five principles since they allow us to compare the (non)explosive behavior of the various logics discussed below in a succinct way. In Section 7 we will consider some additional forms of explosion that can be avoided by using paraconsistent deontic logics.

Benchmark examples. Research in the fields of deontic logic and non-monotonic logic is to a large extent driven by a relatively small set of benchmark examples aimed at testing the formal system in question (the reader may be familiar with Tweety the penguin, the good Samaritan, and the gentle murderer, just to name a few). When faced with such examples, counter-intuitive outcomes are taken to reflect badly on a formal system, so these benchmark examples provide a criterion for checking whether a formal system meets our informal intuitions.

A warning is in order here, however. The fact that a formal system provides intuitive outcomes for the relevant benchmark examples is not a sufficient condition for positively evaluating the system in question. For instance, the system may be devised in an *ad hoc* manner to deal specifically with a small set of examples, at the cost of violating one or more rationality postulates. Moreover, some of these examples may reflect intuitions on which not everyone agrees, leaving room for dispute. In some cases the fact that our logic does *not* give us the expected outcome for some concrete example may inform us that our intuitions are perhaps incoherent, whence this is not in itself a sufficient reason to reject the logic. So, as was the case with explosion principles, we will use our benchmark examples as means to classify given logics, not as absolute criteria for their usefulness.[31]

With this warning in mind, let us list a number of examples which have been used to evaluate conflict-tolerant deontic logics studied in the literature. For each

[31] For a critical discussion of the use of examples as intuition-pumps in the evaluation of logics for defeasible reasoning, see [78].

of them, we indicate some of the basic **SDL**-principles which allow us to infer the conclusion from the given premises. We use (CL) as a generic name for all inferences that are **CL**-valid.

1. *The Smith Argument.* — (Agg), (Inh), (CL)

   (i) Smith ought to fight in the army or perform alternative service to his country ($\mathsf{O}(f \vee s)$).
   (ii) Smith ought not to fight in the army ($\mathsf{O}\neg f$).
   ∴ (iii) Smith ought to perform alternative service to his country ($\mathsf{O}s$).

2. *The Jones Argument.* — (Inh), (CL)

   (i) Jones ought to tell a joke and sing a song ($\mathsf{O}(j \wedge s)$).
   ∴ (ii) Jones ought to tell a joke ($\mathsf{O}j$).

3. *The Roberts Argument, version 1.* — (Inh), (CL)

   (i) Roberts ought to pay federal taxes and register for national service ($\mathsf{O}(t \wedge r)$).
   (ii) Roberts ought not to pay federal taxes but volunteer to help the homeless in his community ($\mathsf{O}(\neg t \wedge v)$).
   ∴ (iii) Roberts ought to register for national service and ought to volunteer to help the homeless ($\mathsf{O}r \wedge \mathsf{O}v$).

4. *The Roberts Argument, version 2.* — (Inh), (CL), (Agg)

   (i) Roberts ought to pay federal taxes and register for national service ($\mathsf{O}(t \wedge r)$).
   (ii) Roberts ought not to pay federal taxes but volunteer to help the homeless in his community ($\mathsf{O}(\neg t \wedge v)$).
   ∴ (iii) Roberts ought to register for national service and volunteer to help the homeless ($\mathsf{O}(r \wedge v)$).

5. *The Thomas Argument.* — (Inh), (Agg), (CL)

(i) Thomas ought to pay federal taxes and either fight in the army or perform alternative service to his country ($\mathsf{O}(t \wedge (f \vee s))$).
(ii) Thomas ought neither to pay federal taxes nor fight in the army ($\mathsf{O}(\neg t \wedge \neg f)$).
∴ (iii) Thomas ought to perform alternative service to his country ($\mathsf{O}s$).

6. *The Natascha Argument, version 1.* — (K) / (Inh), (Agg), (CL)

(i) Natascha ought to take Sarah to the concert ($\mathsf{O}s$).
(ii) Natascha ought to take Martin to the concert ($\mathsf{O}m$).
(iii) It is not the case that Natascha ought to take Sarah *and* Martin to the concert ($\neg\mathsf{O}(s \wedge m)$).
(iv) If she takes Sarah, she ought to buy an extra ticket ($\mathsf{O}(s \supset t)$).
(v) If she takes Martin, she ought to buy an extra ticket ($\mathsf{O}(m \supset t)$).
∴ (vi) Natascha ought to buy an extra ticket ($\mathsf{O}t$).

7. *The Natascha Argument, version 2.* — (K) / (Inh), (Agg), (CL)

(i) Natascha ought to take Sarah to the concert ($\mathsf{O}s$).
(ii) Natascha ought to take Martin to the concert ($\mathsf{O}m$).
(iii) Natascha ought not to take Sarah *and* Martin to the concert ($\mathsf{O}\neg(s \wedge m)$).
(iv) If she takes Sarah, she ought to buy an extra ticket ($\mathsf{O}(s \supset t)$).
(v) If she takes Martin, she ought to buy an extra ticket ($\mathsf{O}(m \supset t)$).
∴ (vi) Natascha ought to buy an extra ticket ($\mathsf{O}t$).

The Smith argument was first presented by Horty [49; 50; 52; 53]; the name “Smith” is due to Goble [43; 42]. The Jones, Roberts, and Thomas arguments are variations on examples from [43; 42]. The Natascha argument is new.

The validity of these arguments is not undisputed. The Jones argument, for instance, which concerns the application of the inheritance principle (Inh), has been called into question [34; 45; 74]. The Natascha argument concerns the derivation of a so-called *floating conclusion*, a conclusion entailed by each of two mutually conflicting obligations. The status of such conclusions is debatable.[32]

[32] See [51; 57; 78] for arguments pro and contra the derivation of floating conclusions in non-monotonic logic. In a moral context, the derivability of floating conclusions has been defended by Brink [26].

In both versions of the Natascha argument, the idea behind the third premise is that for some reason or another, Natascha cannot possibly take both Sarah and Martin to the concert — e.g. because there is only one additional ticket left at the counter. In the absence of alethic modalities, we translate information concerning what is (im)possible directly into the language of **SDL**. While the first version of this argument relies on the principle of "ought implies can" (OIC) and contraposition, the second relies on the stronger principle of "permitted implies can" (PIC), interdefinability of O and P, and contraposition. Both (OIC) and (PIC) are controversial.[33] However, here we focus merely on the formal premises as such, not on the question whether they represent the example in the most natural way.

5 Adaptive inheritance

The first type of conflict-tolerant deontic logics mentioned in Section 4.1 is obtained by giving up or weakening the rule of inheritance (Inh). In the present section, we discuss one specific subclass of such logics, showing how they can be strengthened by going adaptive.

5.1 Logics with unconflicted inheritance

Restricting inheritance In a number of papers, Goble presented the **LUM**-family of deontic logics.[34] The language of these logics is just that of **SDL**, with P defined as the dual of O. The logics in the **LUM**-family do not simply reject inheritance, but replace it with a weaker principle that accounts for a number of intuitive applications of (Inh). This requires some explanation.

Let $\mathsf{U}A =_{\mathsf{df}} \neg(\mathsf{O}A \wedge \mathsf{O}\neg A)$ denote that A is unconflicted. All **LUM**-systems extend **CL** with the necessitation rule (N), the replacement of equivalents rule (RE), as well as the following rule of "unconflicted" inheritance (RUM):

$$\text{If } A \vdash B, \text{ then } \mathsf{U}A, \mathsf{O}A \vdash \mathsf{O}B \qquad \text{(RUM)}$$

(RUM) allows for those applications of the inheritance rule (Inh) which involve only unconflicted obligations. In terms of permission, the rule states that whenever A is both obligatory and permitted, then whatever is logically weaker than A is also obligatory. This rule is therefore also sometimes referred to as "permitted inheritance" (RPM).

[33] See [110] for a comprehensive discussion of the first of these two principles.

[34] We adopt the presentation and nomenclature from [43]. For more details and references, we refer to Section 5.3.

In addition to (N), (RE), and (RUM), the systems in the **LUM**-family are defined in terms of (a selection among) (P), (Agg), and "consistent" and "permitted" aggregation rules (C-Agg) and (P-Agg):

$$\text{If } \nvdash \neg(A \wedge B) \text{ then } \mathsf{O}A, \mathsf{O}B \vdash \mathsf{O}(A \wedge B) \qquad \text{(C-Agg)}$$

$$\mathsf{P}A, \mathsf{P}B, \mathsf{O}A, \mathsf{O}B \vdash \mathsf{O}(A \wedge B) \qquad \text{(P-Agg)}$$

Note that, since P is the dual of O, the antecedent of (P-Agg) just means that A and B are obligatory, and that neither of their negations are obligatory. The systems **LUM.a**-**LUM.c** extend **CL** by adding:

LUM.a: (N), (RE), (RUM), (Agg)
LUM.b: (N), (RE), (RUM), (P), (C-Agg)
LUM.c: (N), (RE), (RUM), (P), (P-Agg)

A semantics for these three logics can easily be obtained, following the well-known generalization of Kripke-semantics into neighbourhood semantics – cf. [29, Chapters 7 & 8] and [85]. Say a **LUM**-model is of the type $M = \langle W, w_0, n_{\mathsf{O}}, v\rangle$, where W is a non-empty set of worlds, $w_0 \in W$ is the actual world, $n_{\mathsf{O}} : W \to \wp(\wp(W))$ maps each world $w \in W$ to the set of *obligatory propositions at* w, and v is a valuation function. The semantic clause for O in such models reads:

(SC-O) $M, w \models \mathsf{O}A$ iff $|A|_M \in n_{\mathsf{O}}(w)$

Truth in a model is defined as usual, viz. as truth at w_0; semantic consequence is defined by quantifying over all models in which the premises are true.

This gives us the minimal classical modal logic **E**, which is characterized fully by adding (RE) to **CL**. Imposing a number of restrictions on such models, we obtain the additional axioms and rules listed above. These conditions are:

(CO-RUM) if $X \in n_{\mathsf{O}}(w)$, $W \setminus X \notin n_{\mathsf{O}}(w)$, and $X \subseteq Y$, then $Y \in n_{\mathsf{O}}(w)$
(CO-N) $W \in n_{\mathsf{O}}(w)$
(CO-P) $\emptyset \notin n_{\mathsf{O}}(w)$
(CO-Agg) if $X \in n_{\mathsf{O}}(w)$ and $Y \in n_{\mathsf{O}}(w)$, then $X \cap Y \in n_{\mathsf{O}}(w)$
(CO-C-Agg) if $X \in n_{\mathsf{O}}(w)$, $Y \in n_{\mathsf{O}}(w)$, and $X \cap Y \neq \emptyset$, then $X \cap Y \in n_{\mathsf{O}}(w)$
(CO-P-Agg) if $X \in n_{\mathsf{O}}(w)$, $Y \in n_{\mathsf{O}}(w)$, $W \setminus X \notin n_{\mathsf{O}}(w)$, and $W \setminus Y \notin n_{\mathsf{O}}(w)$, then $X \cap Y \in n_{\mathsf{O}}(w)$

For an extensive comparison and discussion of the various **LUM**-logics, we refer to [42, Sect. 5.3]. In the remainder, we will focus on ALs obtained from them.

Going adaptive To understand the specific motivation for going adaptive in the case of the **LUM**-logics, it will be useful to reconsider the benchmark examples from Section 4.2. The Smith and Jones arguments are invalid in all three of the **LUM**-logics, but valid once we add the premises $\mathsf{U}\neg f$ (for the Smith argument) and $\mathsf{U}(j \wedge s)$ (for the Jones argument). The Roberts and Thomas arguments are more problematic. In the Roberts argument, for instance, we cannot just add the premises $\mathsf{U}(t \wedge r)$ and $\mathsf{U}(\neg t \wedge v)$ in order to render the argument valid, since doing so would trivialize the premise set.[35]

More generally, it is problematic that in the **LUM**-systems we need to add the 'tacit' information that a formula is unconflicted before we can apply the restricted distribution rule. This worry was first raised in [94], and acknowledged by Goble:

> For one thing, the additional non-conflict condition on the distribution rule seems rather *ad hoc*; there is little to recommend it except its success in disarming deontic explosion. For another, it seems risky to try to account for the plausibility of arguments by considering them enthymematic for straight-forwardly valid arguments. In context it may be all right to accept the alleged tacit premise, but we cannot rely on that. With more complicated arguments it might be quite uncertain what unspoken premises of non-conflict are implicitly present [43, pp. 210-211].

Both problems can be overcome by strengthening the **LUM**-systems within the adaptive logics framework. On the one hand, we can validate all those applications of distribution that do not lead to deontic explosion. On the other hand, it is the logic itself that fixes which applications of distribution are tolerable; no interference of any user is required for this. We explain how this works below, focusing on the adaptive extensions of the logic **LUM.a**. For the other logics in this family, the difficulties and properties are roughly analogous. We will point out salient differences as we go along.

The logics LUM.a$^{\mathbf{x}}$ A natural way of strengthening Goble's **LUM**-systems is to work under the assumption that obligations are unconflicted, so that an obligation $\mathsf{O}A$ behaves abnormally in case it is conflicted, i.e. in case $\neg\mathsf{U}A$ or, equivalently, $\mathsf{O}A \wedge \mathsf{O}\neg A$:

$$\Omega = \{\mathsf{O}A \wedge \mathsf{O}\neg A \mid A \in \mathcal{W}\}$$

[35] For Roberts, first note that $\vdash (t \wedge r) \supset \neg(\neg t \wedge v)$. By (RPM), $\neg\mathsf{O}\neg(t \wedge r) \supset (\mathsf{O}(t \wedge r) \supset \mathsf{O}\neg(\neg t \wedge v))$ or, equivalently, $\mathsf{O}\neg(t \wedge r) \vee (\mathsf{O}(t \wedge r) \supset \mathsf{O}\neg(\neg t \wedge v))$. By premises (i) and (ii) of the Roberts argument, we get $(\mathsf{O}(t \wedge r) \wedge \mathsf{O}\neg(t \wedge r)) \vee (\mathsf{O}(\neg t \wedge v) \wedge \mathsf{O}\neg(\neg t \wedge v))$ by **CL**. So adding $\mathsf{U}(t \wedge r)$ and $\mathsf{U}(\neg t \wedge v)$ would make the argument **CL**-inconsistent. For Thomas the argument is analogous.

The logic $\mathbf{ADPM.1^r}$ from [94] is the AL defined by the triple $\langle \mathbf{LUM.a}, \Omega,$ reliability$\rangle$. In an $\mathbf{ADPM.1^r}$-proof, (Inh) can be applied via the conditional rule RC, assuming that the obligations involved are not conflicted:

1	$\mathsf{O}(p \wedge q)$	Prem	$\emptyset$
2	$\mathsf{O}p$	1; RC	$\{\mathsf{O}(p \wedge q) \wedge \mathsf{O}\neg(p \wedge q)\}$

The conditional derivation at line 2 is legitimate in view of the $\mathbf{LUM.a}$-valid inference

$$\mathsf{O}(p \wedge q) \vdash \mathsf{O}p \vee (\mathsf{O}(p \wedge q) \wedge \mathsf{O}\neg(p \wedge q)) \tag{11}$$

Unfortunately, Goble pointed out that $\mathbf{ADPM.1^r}$ suffers from a problem [43, Sect. 4.3.1]. Although we can indeed apply distribution conditionally in $\mathbf{ADPM.1^r}$, the corresponding application of RC in the proof is marked as soon as a (possibly unrelated) conflict follows from the premise set. The problem is best illustrated by means of a simple example.

1	$\mathsf{O}(p \wedge q)$	Prem	$\emptyset$
2	$\mathsf{O}r$	Prem	$\emptyset$
3	$\mathsf{O}\neg r$	Prem	$\emptyset$
4	$\mathsf{O}p$	1; RC	$\{\mathsf{O}(p \wedge q) \wedge \mathsf{O}\neg(p \wedge q)\}$✓
5	$(\mathsf{O}(p \wedge q) \wedge \mathsf{O}\neg(p \wedge q)) \vee$ $(\mathsf{O}(p \wedge r) \wedge \mathsf{O}\neg(p \wedge r)) \vee$ $(\mathsf{O}(p \wedge \neg r) \wedge \mathsf{O}\neg(p \wedge \neg r))$	1-3; RU	$\emptyset$

The Dab-formula derived at line 5 is minimal at this stage of the proof, and causes the marking of line 4.[36] This Dab-formula is a minimal Dab-consequence of the premise set $\{\mathsf{O}(p \wedge q), \mathsf{O}r, \mathsf{O}\neg r\}$. Consequently, there is no extension of this proof in which line 4 is unmarked, and hence

$$\mathsf{O}(p \wedge q), \mathsf{O}r, \mathsf{O}\neg r \nvdash_{\mathbf{ADPM.1^r}} \mathsf{O}p \tag{12}$$

The same holds if we use the minimal abnormality strategy instead of reliability (the reasoning is analogous):

$$\mathsf{O}(p \wedge q), \mathsf{O}r, \mathsf{O}\neg r \nvdash_{\mathbf{ADPM.1^m}} \mathsf{O}p \tag{13}$$

This problem generalizes: in the presence of a conflict between two obligations, we can construct minimal Dab-formulas containing abnormalities pertaining to seemingly unrelated and unproblematic formulas, blocking unproblematic applications

[36] By (11), $\mathsf{O}p \vee (\mathsf{O}(p \wedge q \wedge \mathsf{O}\neg(p \wedge q))$. Suppose $\mathsf{O}p$. Then (i) by (Agg), $\mathsf{O}(p \wedge r)$ and, by (RUM) and $\mathbf{CL}$, $\mathsf{O}\neg(p \wedge \neg r) \vee (\mathsf{O}(p \wedge r) \wedge \mathsf{O}\neg(p \wedge r))$; analogously (ii) by (Agg), $\mathsf{O}(p \wedge \neg r)$ and, by (RUM) and $\mathbf{CL}$, $\mathsf{O}\neg(p \wedge r) \vee (\mathsf{O}(p \wedge \neg r) \wedge \mathsf{O}\neg(p \wedge \neg r))$. Altogether, by $\mathbf{CL}$, $(\mathsf{O}(p \wedge q) \wedge \mathsf{O}\neg(p \wedge q)) \vee (\mathsf{O}(p \wedge r) \wedge \mathsf{O}\neg(p \wedge r)) \vee (\mathsf{O}(p \wedge \neg r) \wedge \mathsf{O}\neg(p \wedge \neg r))$.

of RC. The logics $\mathbf{ADPM.1^r}$ and $\mathbf{ADPM.1^m}$ are therefore called *flip-flops* [10]. In the absence of conflicts, their consequence set is the same as their ULL, namely **SDL**.[37] As soon as one conflict is present, however, their consequence set collapses into that of their lower limit logic **LUM.a**.

There is a natural fix to this flip-flop problem, due to Goble [43]. Let $\mathrm{S}(A)$ denote the set of all subformulas of A (including A itself). Where $\mathrm{S}(A) = \{B_1, \ldots, B_n\}$, we define[38]

$$\sharp(A) = (\mathsf{O}B_1 \wedge \mathsf{O}\neg B_1) \vee \ldots \vee (\mathsf{O}B_n \wedge \mathsf{O}\neg B_n)$$

Following Goble, we let $\mathbf{LUM.a^r} = \langle \mathbf{LUM.a}, \Omega^{\mathrm{S}}, \text{reliability}\rangle$, where[39]

$$\Omega^{\mathrm{S}} = \{\sharp(A) \mid A \in \mathcal{W}\}$$

In an $\mathbf{LUM.a^r}$-proof, the formula derived at line 5 of our proof above is no longer a Dab-formula. Rather, we obtain the following proof:

1	$\mathsf{O}(p \wedge q)$	Prem	$\emptyset$
2	$\mathsf{O}r$	Prem	$\emptyset$
3	$\mathsf{O}\neg r$	Prem	$\emptyset$
4	$\mathsf{O}p$	1; RC	$\{\sharp(p \wedge q)\}$
5	$\sharp(p \wedge q) \vee \sharp(p \wedge r) \vee \sharp(p \wedge \neg r)$	1-3; RU	$\emptyset$
6	$\sharp(p \wedge r)$	2,3; RU	$\emptyset$
7	$\sharp(p \wedge \neg r)$	2,3; RU	$\emptyset$

The abnormalities $\sharp(p \wedge q)$, $\sharp(p \wedge r)$, and $\sharp(p \wedge \neg r)$ denote the formulas (14), (15), and (16) respectively:

$$(\mathsf{O}(p \wedge q) \wedge \mathsf{O}\neg(p \wedge q)) \vee (\mathsf{O}p \wedge \mathsf{O}\neg p) \vee (\mathsf{O}q \wedge \mathsf{O}\neg q) \tag{14}$$

$$(\mathsf{O}(p \wedge r) \wedge \mathsf{O}\neg(p \wedge r)) \vee (\mathsf{O}p \wedge \mathsf{O}\neg p) \vee (\mathsf{O}r \wedge \mathsf{O}\neg r) \tag{15}$$

$$(\mathsf{O}(p \wedge \neg r) \wedge \mathsf{O}\neg(p \wedge \neg r)) \vee (\mathsf{O}p \wedge \mathsf{O}\neg p) \vee (\mathsf{O}\neg r \wedge \mathsf{O}\neg\neg r) \vee (\mathsf{O}r \wedge \mathsf{O}\neg r) \tag{16}$$

The inference made at line 4 is legitimate in view of the **LUM.a**-valid inference

$$\mathsf{O}(p \wedge q) \vdash \mathsf{O}p \vee \sharp(p \wedge q) \tag{17}$$

Since $\sharp(p \wedge r)$ and $\sharp(p \wedge \neg r)$ are **LUM.a**-derivable from the premises $\mathsf{O}r$ and $\mathsf{O}\neg r$, the Dab-formula derived at line 5 of the proof is not minimal at stage 7. Consequently,

[37] It was shown in [94, Th. 7] that **SDL** is the ULL of $\mathbf{ADPM.1^r}$.

[38] Our expression $\sharp(A)$ is equivalent to the negation of Goble's expression $\mho(A)$ in [43; 42]. Note that $\sharp$ is not a (modal or other) operator but just a symbol that allows us to abbreviate a formula.

[39] Goble uses the name $\mathbf{ALUM^r}$ for the logic that we call $\mathbf{LUM.a^r}$.

line 4 is unmarked at this stage. As opposed to $\mathbf{ADPM.1^r}$ and $\mathbf{ADPM.1^m}$, the logics $\mathbf{LUM.a^r}$ and $\mathbf{LUM.a^m}$ lead to the following desirable outcome:

$$\mathsf{O}(p \wedge q), \mathsf{O}r, \mathsf{O}\neg r \vdash_{\mathbf{LUM.a^r}} \mathsf{O}p \tag{18}$$

$$\mathsf{O}(p \wedge q), \mathsf{O}r, \mathsf{O}\neg r \vdash_{\mathbf{LUM.a^m}} \mathsf{O}p \tag{19}$$

5.2 Evaluating the logics

Explosion principles The adaptive logics based on the **LUM**-family are conflict-tolerant to the same extent as their respective lower limit logics. This means, for a start, that (DEX) is invalid in all of them. Since they are **CL**-based and in view of the interdefinability of O and P, they also accommodate conflicts of the form $\mathsf{O}A \wedge \neg\mathsf{P}A$, which simply reduce to conflicts between obligations.

However, the logics do not tolerate the other types of deontic conflicts that were discussed in Section 4.2. While $\mathsf{O}(A \wedge \neg A)$ is consistent in **LUM.a** – and hence also in $\mathbf{LUM.a^x}$, it is inconsistent in each of **LUM.b** and **LUM.c** in view of the (P)-axiom. It follows that ALs based on the latter two logics cannot make sense of self-contradictory obligations. Also, all the (adaptive) **LUM**-logics trivialize conflicts of the form $\mathsf{O}A \wedge \mathsf{P}\neg A$, as these reduce to plain contradictions in view of (Def_{P}) and (RE). Finally, $\mathsf{P}(A \wedge \neg A)$ (which is equivalent to $\neg\mathsf{O}(\neg A \vee A)$) is also trivial in these logics, in view of the necessitation rule (N).

Benchmark examples The Smith and Jones arguments are $\mathbf{LUM.a^x}$-valid. The premises of these arguments are **SDL**-consistent and hence normal, which means that (by Theorem 3.14), the adaptive logics are just as strong as **SDL** for these cases.[40] The Roberts and Thomas arguments are not valid in $\mathbf{LUM.a^r}$ or $\mathbf{LUM.a^m}$. Here is a proof illustrating why the Roberts arguments are not valid in $\mathbf{LUM.a^x}$:

1	$\mathsf{O}(t \wedge r)$	Prem	$\emptyset$
2	$\mathsf{O}(\neg t \wedge v)$	Prem	$\emptyset$
3	$\mathsf{O}r$	1; RC	$\{\sharp(t \wedge r)\}\checkmark$
4	$\mathsf{O}v$	2; RC	$\{\sharp(\neg t \wedge v)\}\checkmark$
5	$\mathsf{O}(r \wedge v)$	3,4; RU	$\{\sharp(t \wedge r), \sharp(\neg t \wedge v)\}\checkmark$
6	$\sharp(t \wedge r) \vee \sharp(\neg t \wedge v)$	1,2; RU	$\emptyset$

In order to infer $\mathsf{O}r$ and $\mathsf{O}v$ via RC we need to rely on the falsity of $\sharp(t \wedge r)$ and $\sharp(\neg t \wedge v)$. However, further inspection of the premises teaches us that the disjunction of these abnormalities is **LUM.a**-derivable from the premises. To see

[40] Goble showed that the upper limit logic of $\mathbf{LUM.a^x}$ is **SDL**, see [43, Observation 4.1].

why, note that this disjunction is **LUM.a**-equivalent to the following formula, which is a **LUM.a**-consequence of the premises:[41]

$$\begin{aligned}(\mathsf{O}(t \wedge r) \wedge \mathsf{O}\neg(t \wedge r)) \vee (\mathsf{O}(\neg t \wedge v) \wedge \mathsf{O}\neg(\neg t \wedge v)) \vee \\ (\mathsf{O}t \wedge \mathsf{O}\neg t) \vee (\mathsf{O}r \wedge \mathsf{O}\neg r) \vee (\mathsf{O}v \wedge \mathsf{O}\neg v)\end{aligned} \tag{20}$$

The minimal Dab-formula derived at line 6 blocks the derivation of the formulas derived at lines 3-5, causing the invalidity of the Roberts arguments. The same mechanism blocks the derivation of the conclusion of the Thomas argument.[42]

The Natascha argument, version 1, is **LUM.a**-valid (and hence $\mathbf{LUM.a^x}$-valid), but only because its premise set is **LUM.a**-trivial: from premises (i) and (ii) we can derive the negation of premise (iii) by (Agg). In contrast, the second version of the Natascha argument is **LUM.a**-satisfiable. Here is an $\mathbf{LUM.a^m}$-proof for this argument:

1	$\mathsf{O}s$	Prem	$\emptyset$
2	$\mathsf{O}m$	Prem	$\emptyset$
3	$\mathsf{O}\neg(s \wedge m)$	Prem	$\emptyset$
4	$\mathsf{O}(s \supset t)$	Prem	$\emptyset$
5	$\mathsf{O}(m \supset t)$	Prem	$\emptyset$
6	$\mathsf{O}(s \wedge t)$	1, 4; RU	$\emptyset$
7	$\mathsf{O}(m \wedge t)$	2, 5; RU	$\emptyset$
8	$\mathsf{O}t$	6; RC	$\{\sharp(s \wedge t)\}$
9	$\mathsf{O}t$	7; RC	$\{\sharp(m \wedge t)\}$
10	$\sharp(s \wedge t) \vee \sharp(m \wedge t)$	1-3; RU	$\emptyset$

The formulas derived at lines 6 and 7 are **LUM.a**-derivable from the premises via applications of (Agg) and (RE). From each of these formulas we can derive $\mathsf{O}t$ via RC. Since we are working with the minimal abnormality strategy, lines 8 and 9 are unmarked at stage 10. If we were to use reliability, however, both lines would be marked. Indeed, the modified Natascha argument is valid for $\mathbf{LUM.a^m}$, while invalid for $\mathbf{LUM.a^r}$:

$$\mathsf{O}s, \mathsf{O}m, \mathsf{O}\neg(s \wedge m), \mathsf{O}(s \supset t), \mathsf{O}(m \supset t) \nvdash_{\mathbf{LUM.a^r}} \mathsf{O}t \tag{21}$$

$$\mathsf{O}s, \mathsf{O}m, \mathsf{O}\neg(s \wedge m), \mathsf{O}(s \supset t), \mathsf{O}(m \supset t) \vdash_{\mathbf{LUM.a^m}} \mathsf{O}t \tag{22}$$

[41] By **CL**, $\big(\mathsf{O}(t \wedge r) \wedge \mathsf{O}\neg(t \wedge r)\big) \vee \neg\big(\mathsf{O}(t \wedge r) \wedge \mathsf{O}\neg(t \wedge r)\big)$. Since $\mathsf{O}(t \wedge r)$, $\neg\big(\mathsf{O}(t \wedge r) \wedge \mathsf{O}\neg(t \wedge r)\big)$ entails $\mathsf{O}t$ by (RUM). Analogously, by **CL**, $\big(\mathsf{O}(\neg t \wedge v) \wedge \mathsf{O}\neg(\neg t \wedge v)\big) \vee \neg\big(\mathsf{O}(\neg t \wedge v) \wedge \mathsf{O}\neg(\neg t \wedge v)\big)$. Since $\mathsf{O}(\neg t \wedge v)$, $\neg\big(\mathsf{O}(\neg t \wedge v) \wedge \mathsf{O}\neg(\neg t \wedge v)\big)$ entails $\mathsf{O}\neg t$ by (RUM). Altogether, by **CL**, $\big(\mathsf{O}(t \wedge r) \wedge \mathsf{O}\neg(t \wedge r)\big) \vee \big(\mathsf{O}(\neg t \wedge v) \wedge \mathsf{O}\neg(\neg t \wedge v)\big) \vee (\mathsf{O}t \wedge \mathsf{O}\neg t)$. By **CL** again, (20) follows.

[42] In the Thomas case, the culpable Dab-formula is the disjunction $\sharp(t \wedge (f \vee s)) \vee \sharp(\neg t \wedge \neg f)$. We leave the verification to the interested reader.

The behavior of the ALs based on **LUM.b** and **LUM.c** is roughly analogous to the preceding case, with one exception. The premises in version 1 of the Natascha argument are inconsistent in **LUM.a** and **LUM.b**, but consistent in **LUM.c**. That is, we cannot aggregate premises (i) and (ii) of this argument, in the absence of the permission statements $\mathsf{P}s$ and $\mathsf{P}m$. Parallel to the situation for the modified Natascha argument in $\mathbf{LUM.a^x}$, we obtain the conclusion $\mathsf{O}t$ with $\mathbf{LUM.c^m}$ for the original Natascha argument, while we do not obtain it with $\mathbf{LUM.c^r}$. The following proof illustrates that $\mathsf{O}t$ is $\mathbf{LUM.c^m}$-derivable:[43]

1	$\mathsf{O}s$	Prem	$\emptyset$
2	$\mathsf{O}m$	Prem	$\emptyset$
3	$\neg\mathsf{O}(s \wedge m)$	Prem	$\emptyset$
4	$\mathsf{O}(s \supset t)$	Prem	$\emptyset$
5	$\mathsf{O}(m \supset t)$	Prem	$\emptyset$
6	$\mathsf{O}(s \wedge m)$	1,2; RC	$\{\sharp s, \sharp m\}\checkmark^{12}$
7	$\mathsf{O}(s \wedge t)$	1, 4; RU	$\{\sharp s, \sharp(s \supset t)\}\checkmark^{12}$
8	$\mathsf{O}(m \wedge t)$	2, 5; RU	$\{\sharp m, \sharp(m \supset t)\}\checkmark^{12}$
9	$\mathsf{O}t$	7; RC	$\{\sharp s, \sharp(s \supset t), \sharp(s \wedge t)\}$
10	$\mathsf{O}t$	8; RC	$\{\sharp m, \sharp(m \supset t), \sharp(m \wedge t)\}$
11	$\sharp(s \wedge t) \vee \sharp(m \wedge t)$	1-3; RU	$\emptyset$
12	$\sharp s \vee \sharp m$	1-3; RU	$\emptyset$
13	$\sharp s \vee \sharp(m \wedge t)$	12; RU	$\emptyset$
14	$\sharp(s \wedge t) \vee \sharp m$	12; RU	$\emptyset$

The inferences at lines 13 and 14 hold in view of the **CL**-validity of $\sharp s \supset \sharp(s \wedge t)$ and $\sharp m \supset \sharp(m \wedge t)$ respectively. Where $\Gamma_n = \{\mathsf{O}s, \mathsf{O}m, \neg\mathsf{O}(s \wedge m), \mathsf{O}(s \supset t), \mathsf{O}(m \supset t)\}$:

$$\Phi_{14}(\Gamma_n) = \{\{\sharp(s \wedge t), \sharp s\}, \{\sharp(m \wedge t), \sharp m\}\} \tag{23}$$

It is easily verified that, in view of Definition 3.2, lines 9 and 10 are unmarked. If we were to use the reliability strategy instead, then by Definition 3.1 these lines would be marked in the proof above.

$$\Gamma_n \nvdash_{\mathbf{LUM.c^r}} \mathsf{O}t \tag{24}$$

$$\Gamma_n \vdash_{\mathbf{LUM.c^m}} \mathsf{O}t \tag{25}$$

The formula $\mathsf{O}t$ is a floating conclusion with respect to Γ_n. As pointed out in Section 4, it is a matter of debate whether or not floating conclusions are acceptable.

[43] Lines 7 and 8 can be derived by means of (P-Agg). Note that this rule requires that the two formulas to be aggregated are themselves unconflicted. Hence we need RC to make these two derivations.

We do not add anything to this debate here. It suffices for us to point out that each stance can be formally represented within the AL framework.

5.3 Further reading and open ends

The **LUM**-systems were introduced by Goble in [35; 41; 36], where they were called 'logics of permitted distribution' or **DPM**. They were called 'logics of unconflicted distribution' or **LUM** in [42; 43]. Adaptive extensions of these systems were presented in [94; 87; 43]. Moreover, in [91], dyadic variants of the **LUM**-systems were also strengthened within the AL framework (see also Section 9.1 below).

There are many other types of deontic logics which invalidate (Inh). First, there is the general class of classical modal logics of which the logic **E** (cf. supra) is but one example. Second, Goble [34; 37] developed a very rich semantics for deontic logics, based on an idea from [54]. On this semantics, $\mathsf{O}A$ is true iff the closest A-worlds are all better than the closest $\neg A$-worlds. Third and last, in more recent work, Cariani [27] proposed yet another semantics for "ought" which invalidates (Inh) in a principled way – see also [100; 101] for a formal investigation into this proposal. For each of these types of logics, one can ask whether it makes sense to strengthen them adaptively, and if so, which technical difficulties arise and what behavior the resulting logics will display. In particular, it would be interesting to learn whether some such variants perform better than the currently available logics, in dealing with the Roberts arguments and the Thomas argument.

6 Adaptive aggregation

A popular way to accommodate deontic conflicts in a formal system is by rejecting the aggregation principle (Agg), and with it the normality schema (K). In its simplest form, this proposal gives us the deontic logic **P**.[44] We will focus on two relatively basic ALs obtained from **P** in this section.

6.1 Adaptive aggregation: a basic example

Rejecting aggregation The language of **P** is the same as that of **SDL**, with P defined as the dual of O. As before, we will not consider nested occurrences of O. **P** is axiomatized by adding the axiom (D) to **CL** and closing the resulting set under modus ponens (MP), the necessitation rule (N), and the rule of inheritance (Inh). Each of the following are facts about the derivability relation of **P**:

[44] Again, we follow Goble's nomenclature. See the end of this section for pointers to the literature on this and related logics.

$$\vdash \mathsf{O}(p \vee \neg p) \tag{26}$$

$$\mathsf{O}p \vdash \mathsf{O}(p \vee q) \tag{27}$$

$$\mathsf{O}(p \wedge q) \vdash \mathsf{O}p, \mathsf{O}q \tag{28}$$

$$\mathsf{O}p, \mathsf{O}q \nvdash \mathsf{O}(p \wedge q) \tag{29}$$

$$\mathsf{O}(p \wedge (\neg p \vee q)) \vdash \mathsf{O}q \tag{30}$$

$$\mathsf{O}p, \mathsf{O}(\neg p \vee q) \nvdash \mathsf{O}q \tag{31}$$

In view of (Inh), Replacement of (Classical) Equivalents (RE) is valid in **P**. So in Chellas' terms, **P** is a non-normal but classical modal logic [29].

One way to motivate and understand the rejection of (Agg) in **P** is in terms of multiple normative standards that ground our obligations, where $\mathsf{O}A$ is unspecific about the normative standard that grounds the obligation that A. Under such a reading, $\mathsf{O}A$ and $\mathsf{O}B$ may well be true even if there is no single standard that grounds the conjunction of both obligations, and hence $\mathsf{O}(A \wedge B)$ can still fail.[45] For instance, varying on our Smith example, one's duty to fight in the army might be based on the laws of one's country, whereas one's personal pacifist ethics grounds the claim that one ought not to fight in the army. Still, it does not follow that one ought to do the logically impossible, viz. to fight in the army and not fight in the army.

A semantics for **P** is obtained from the **SDL**-semantics (cf. Section 2) by generalizing the notion of an accessibility relation R. **P**-models are then of the type $\langle W, w_0, \mathcal{R}, V\rangle$, where W, w_0, and V are as before, but $\mathcal{R}$ is a non-empty *set* of serial accessibility relations, rather than a single such relation. The semantic clause for O then reads as follows:

(SC-O) $M, w \models \mathsf{O}A$ iff there is an $R \in \mathcal{R}$ such that $M, w' \models A$ for all w' such that Rww'

In other words, the single normative standard from **SDL** is replaced with a set of such standards, and we quantify (existentially) over such standards in order to determine the truth of $\mathsf{O}A$. It is well-known that **P** is sound and complete with respect to this semantics – see [38, Theorem 1]. Other semantics can also be given for **P**. We refer the reader to [42, pp. 300-301] for an overview of these.

[45] The idea that one can relativize deontic logic to a given "moral code", and that what is obligatory under one such code may not be obligatory (or even forbidden) under another, is at least as old as Von Wright's *Deontic Logic* – see [109, p. 15]. The difference here is that in **P**, the code that is at stake remains implicit, and $\mathsf{O}A$ only means that A is obligatory under at least *some* moral code.

Going adaptive Even if aggregation is invalid on the reading of O just presented, in practice we do often aggregate our obligations. One simple way to argue for this is by referring to the benchmark examples from Section 4. It can easily be verified that neither the Smith argument nor the second variant of the Roberts argument is valid in **P**.

More generally, it is one thing to say that we take into account various normative standards and treat them as independent grounds or reasons when trying to determine what our obligations are. It is quite another thing to argue that none of these obligations can themselves be aggregated when doing so; this seems to go against much of our intuition.[46] For instance, when deciding how to get to the office in the morning, I may apply norms concerning the environment, norms uttered by my boss, and norms concerning my own safety and that of others. There seems to be no *prima facie* reason why we cannot integrate these various norms when settling for a single way to get to the office – e.g. I may conclude that I ought to bike to the office, since that way I will be in time for a meeting without causing air-pollution. The presence of deontic conflicts in itself seems insufficient to warrant a full rejection of aggregation, and, as we will show below, there is no logical reason for doing so either.

One needs to be careful here though. We cannot just add (Agg) to **P**, as this would give us again full **SDL** and hence deontic explosion in the face of deontic conflicts.[47] Moreover, as shown in [41, Sect. 2], there is no obvious conditional variant of (Agg) that can do a similar job, without in turn yielding some variant of deontic explosion.[48] So some obligations can be aggregated, but not all. As we will show in the remainder of this section, going adaptive allows us to steer a middle course between the weakness of **P** and deontic explosion.

The logics $\mathbf{P^x}$ The most straightforward way one might strengthen **P** adaptively, is by treating all formulas of the form $\mathsf{O}A \wedge \mathsf{O}B \wedge \neg\mathsf{O}(A \wedge B)$ as abnormalities. However, just as in the case of $\mathbf{ADPM1^r}$, this will give us a flip-flop. To see why, consider $\Gamma = \{\mathsf{O}p, \mathsf{O}\neg p, \mathsf{O}q, \mathsf{O}r\}$. Intuitively speaking, there is no problem with q and r in this example, and hence we expect $\mathsf{O}(q \wedge r)$ to be derivable. Such an inference can indeed be made within a proof of the adaptive logic thus defined.

[46] Compare [42, p. 253]: "Even if what one ought to do is often determined by different sources or authorities, insofar as propositions of what one ought to do serve as guides to action or as standards of evaluation of an agent's overall actions, there must be a common ought derived from those separate sources".

[47] We safely leave it to the reader to check that adding (Agg) to **P** yields full **SDL**.

[48] See also [42, Section 5.2]. In particular, Goble shows that adding the axiom (C-Agg) (cf. Section 5.1) to **P** will result in a variant of deontic explosion.

However, we can derive a disjunction of abnormalities (in that adaptive logic) from Γ which will block the derivation. This Dab-formula is a disjunction of the following three formulas:

$$\mathsf{O}q \wedge \mathsf{O}r \wedge \neg\mathsf{O}(q \wedge r)) \tag{32}$$

$$\mathsf{O}(p \vee \neg(q \wedge r)) \wedge \mathsf{O}\neg p \wedge \neg\mathsf{O}((p \vee \neg(q \wedge r)) \wedge \neg p) \tag{33}$$

$$\mathsf{O}(q \wedge r) \wedge \mathsf{O}\neg(q \wedge r) \wedge \neg\mathsf{O}((q \wedge r) \wedge \neg(q \wedge r)) \tag{34}$$

Suppose that (32) is false but the premises are true. Then $\mathsf{O}(q \wedge r)$ is the case. Likewise, since $\mathsf{O}(p \vee \neg(q \wedge r))$ follows by (Inh) from $\mathsf{O}p$, (33) can only be false (in view of the premises) if its last conjunct is false, and hence $\mathsf{O}\neg(q \wedge r)$ is true. But then the third abnormality, (34) must be true.

It is not hard to see where the problem could be in cases like this. That is, since $\mathsf{O}p, \mathsf{O}\neg p \in \Gamma$, we should not use these obligations – nor weakenings of them – in order to apply aggregation. In other words, obligations that are themselves conflicted, or subformulas of which are conflicted, should be treated as abnormal.

This brings us to a slightly more complicated set of abnormalities, which is due to Goble [43]. As before, let $\sharp(A)$ denote the disjunction of all formulas $\mathsf{O}B \wedge \mathsf{O}\neg B$, where $B \in S(A)$ (B is a subformula of A). Let $\natural(A, B) = (\mathsf{O}A \wedge \mathsf{O}B \wedge \neg\mathsf{O}(A \wedge B)) \vee \sharp(A \wedge B)$. We now define

$$\Omega_{\mathbf{P}} = \{(\natural(A, B) \mid A, B \in \mathcal{W}\}$$

In other words, we have an abnormality with respect to A and B iff they are both obligatory and their conjunction is not obligatory, *or* a proper subformula of them is conflicted. This means that as soon as e.g. $\mathsf{O}p, \mathsf{O}\neg p$ holds, all abnormalities $\natural(A, B)$ with $p \in S(A)$ are true. Under this definition, none of the formulas (32)-(34) are abnormalities. The corresponding disjunction of $\Omega_{\mathbf{P}}$-abnormalities

$$\natural(q, r) \vee \natural(p \vee \neg(q \wedge r), \neg p) \vee \natural(q \wedge r, \neg(q \wedge r)) \tag{35}$$

is not a minimal Dab-consequence of Γ, since $\natural(p \vee \neg(q \wedge r), \neg p)$ alone follows from Γ.

Let the logics $\mathbf{P^r}$ and $\mathbf{P^m}$ be the adaptive logics defined by the triple $\langle \mathbf{P}, \Omega_{\mathbf{P}}, \mathsf{x}\rangle$, where $\mathsf{x} \in \{\mathsf{r}, \mathsf{m}\}$.[49] It can easily be checked that the upper limit logic of $\mathbf{P^r}$ and $\mathbf{P^m}$ is just **SDL**: adding the negation of all members of $\Omega_{\mathbf{P}}$ as axioms to $\mathbf{P}$, is equivalent

[49] In Goble's work, the first of these two logics is known as $\mathbf{AP^r}$. As before, we skip the initial "A" since the superscript suffices to mark the difference with the monotonic logic $\mathbf{P}$.

to adding (Agg) to **P**.[50] This means that normal premise sets in the logics $\mathbf{P^x}$ are just **SDL**-consistent premise sets (where 'normal' is understood in the technical sense specified on page 23). Hence by Theorem 3.14, whenever a premise set is **SDL**-consistent, its $\mathbf{P^x}$-consequence set will be identical to its **SDL**-consequence set:

Theorem 6.1. *If* Γ *is* **SDL**-*consistent, then* $\mathsf{Cn}_{\mathbf{P^r}}(\Gamma) = \mathsf{Cn}_{\mathbf{P^m}}(\Gamma) = \mathsf{Cn}_{\mathbf{SDL}}(\Gamma)$.

6.2 Evaluating the logics

Explosion principles The logic **P**, and with it $\mathbf{P^r}$ and $\mathbf{P^m}$, clearly accommodates conflicts of the basic type $\mathsf{O}A, \mathsf{O}\neg A$. By (RE), (Def$_\mathsf{P}$) and **CL**-properties, also conflicts of the type $\mathsf{O}A, \neg\mathsf{P}A$ are consistent in **P** and its adaptive extensions.

All other types of deontic conflicts listed in Section 4.2 will be trivialized within these logics. The reasons are similar to those for **LUM.b** and **LUM.c**: $\mathsf{O}(A \wedge \neg A)$ is contradictory in view of (P), $\mathsf{O}A \wedge \mathsf{P}\neg A$ is contradictory in view of (Def$_\mathsf{P}$) and (RE), and $\mathsf{P}(A \wedge \neg A)$ is false in view of (N) and (Def$_\mathsf{P}$). So the simplicity of **P** comes at an important price, viz. that it can only handle conflicting obligations and does not allow us to reason about conflicting information concerning (obligations and) permissions.[51]

Benchmark examples The arguments for Jones and Roberts 1 are valid in both $\mathbf{P^r}$ and $\mathbf{P^m}$. This is easy to verify since the arguments are already valid in **P** in view of its validating (Inh), and since both $\mathbf{P^r}$ and $\mathbf{P^m}$ are extensions of **P**. The premises of the Smith argument are normal: no Dab-formula can be derived from them. As a result, we can aggregate the obligations in the argument and derive $\mathsf{O}s$.

The second Roberts argument is also valid in $\mathbf{P^x}$, but here the reasoning is slightly more intricate. First, applying (Inh), we can derive $\mathsf{O}r$ and $\mathsf{O}v$ from the premises. To apply aggregation to these two formulas, we need to assume that neither r nor v are conflicted, given the premise set. This is clearly the case: the only conflict that follows from the premises, is $\mathsf{O}t, \mathsf{O}\neg t$. The following $\mathbf{P^r}$-proof illustrates how we can obtain the desired conclusion for Roberts 2, while avoiding the aggregation of conflicted obligations:

[50] To see why this is so, note first that if we negate all formulas of the form $\natural(A, B)$, then *a fortiori* we negate all formulas of the form $\mathsf{O}A \wedge \mathsf{O}B \wedge \neg\mathsf{O}(A \wedge B)$, and hence we affirm all instances of (Agg). In addition, we also negate all formulas of the form $\mathsf{O}A \wedge \mathsf{O}\neg A$, but these are anyway **SDL**-valid.

[51] Note also that simply rejecting (P) will not allow us to have a satisfactory account of conflicts of the type $\mathsf{O}(A \wedge \neg A)$: due to (Inh) these conflicts will still lead to deontic explosion.

1	$\mathsf{O}(t \wedge r)$	Prem	$\emptyset$
2	$\mathsf{O}(\neg t \wedge v)$	Prem	$\emptyset$
3	$\mathsf{O}r$	1; RU	$\emptyset$
4	$\mathsf{O}v$	2; RU	$\emptyset$
5	$\mathsf{O}(r \wedge v)$	4,5; RC	$\{\natural(r, v)\}$
6	$\mathsf{O}t$	1; RU	$\emptyset$
7	$\mathsf{O}\neg t$	2; RU	$\emptyset$
8	$\mathsf{O}(t \wedge v)$	6,4; RU	$\{\natural(t, v)\}\checkmark^{11}$
9	$\mathsf{O}(t \wedge \neg t)$	6,7; RC	$\{\natural(t, \neg t)\}\checkmark^{10}$
10	$\natural(t, \neg t)$	6,7; RU	$\emptyset$
11	$\natural(t, v)$	6,7; RU	$\emptyset$

Since $\natural(t, v)$ follows from the premises, we cannot finally derive $\mathsf{O}(t \wedge v)$ from them. So even if there is no direct conflict between t and v, the fact that t is itself conflicted is sufficient to block its aggregation with other (unproblematic) obligations.[52]

The reasoning for the Thomas argument is wholly analogous to the second Roberts case, with the difference that we apply (Inh) once more after aggregating $\mathsf{O}(f \vee s)$ and $\mathsf{O}\neg f$ to $\mathsf{O}((f \vee s) \wedge \neg s)$. This gives us the desired conclusion $\mathsf{O}s$.

For the Natascha arguments, it turns out that with the **P**-based adaptive logics the strategies make no difference. The point is that, although we can obviously not apply aggregation to $\mathsf{O}s$ and $\mathsf{O}m$, we can still aggregate $\mathsf{O}s$ and $\mathsf{O}(s \supset t)$ (and likewise, $\mathsf{O}m$ and $\mathsf{O}(m \supset t)$). The fact that the pair (m, s) behaves abnormally ($\natural(m, s)$ follows from the premises of the argument) does not imply that either of $(s, s \supset t)$ or $(m, m \supset t)$ behave abnormally. Hence we can finally derive $\mathsf{O}t$ on two different conditions in both $\mathbf{P^r}$ and $\mathbf{P^m}$. We illustrate this for the first variant of the Natascha argument:

[52] As pointed out by Goble, allowing aggregation for all A, B such that $A \wedge B$ is consistent is simply a no-go in the context of **P**, since it will lead to another form of deontic explosion. See [41, Sect. 2.4.1].

1	$\mathsf{O}s$	Prem	$\emptyset$
2	$\mathsf{O}m$	Prem	$\emptyset$
3	$\neg\mathsf{O}(s \wedge m)$	Prem	$\emptyset$
4	$\mathsf{O}(s \supset t)$	Prem	$\emptyset$
5	$\mathsf{O}(m \supset t)$	Prem	$\emptyset$
6	$\mathsf{O}(s \wedge (s \supset t))$	1,4; RC	$\{\natural(s, s \supset t)\}$
7	$\mathsf{O}(m \wedge (m \supset t))$	2, 5; RC	$\{\natural(m, m \supset t)\}$
8	$\mathsf{O}t$	6; RU	$\{\natural(s, s \supset t)\}$
9	$\mathsf{O}t$	7; RU	$\{\natural(m, m \supset t)\}$

For none of the variants of the Natascha arguments, the disjunction of abnormalities $(\mathsf{O}s \wedge \mathsf{O}\neg s) \vee (\mathsf{O}m \wedge \mathsf{O}\neg m)$ is **P**-derivable from the premises. Nor is there another Dab-formula which prevents lines 8 and 9 from being finally derivable. So, to sum up, all inferences from our benchmark examples are valid in the logics $\mathbf{P^r}$ and $\mathbf{P^m}$.

This is not to say that there is no difference between $\mathbf{P^r}$ and $\mathbf{P^m}$. Consider e.g. $\Gamma = \{\mathsf{O}p, \mathsf{O}q, \mathsf{O}r, \neg\mathsf{O}(p \wedge r) \vee \neg\mathsf{O}(q \wedge r)\}$. From this premise set, $\mathbf{P^r}$ will not allow us to finally derive $\mathsf{O}(p \wedge r) \vee \mathsf{O}(q \wedge r)$, whereas $\mathbf{P^m}$ will. To understand this, note that $\natural(p, r) \vee \natural(q, r)$ is a minimal Dab-consequence of Γ, whence both abnormalities are unreliable in view of Γ. However, since nothing prevents us to assume that either the first or the second abnormality is *false*, using *minimal abnormality* we can derive $\mathsf{O}(p \wedge r) \vee \mathsf{O}(q \wedge r)$.

6.3 Further reading and open ends

Bernard Williams [111] was the first to advocate a rejection of (Agg) on philosophical grounds; Marcus [62] is another important proponent of such a rejection. More formally worked out proposals can be found in [107; 29; 84]. Later, Goble developed the semantics and metatheory of **P** and variants of it in detail – see in particular [38; 40; 39]. For a more complete overview of the literature on **P** and close (monotonic) relatives, we refer to [42, Section 5.2].

The first adaptive logic that applies the idea of “adaptive aggregation” was published in [66], and later reworked in [67]. These logics are however based on a richer lower limit logic, viz. the logic $\mathbf{SDL}_a\mathbf{P}_e$ from [38]. In this system, one can express both an “existential” notion of obligation O_e (whose logic is **P**) and a “universal” notion of obligation O_a, whose logic is **SDL**. The two modalities are connected by the following bridging principle:

(B) $\mathsf{O}_a(A \supset B) \wedge \mathsf{O}_e A \vdash \mathsf{O}_e B$

which entails i.a. that every universal obligation is also an existential obligation, $\mathsf{O}_a A \supset \mathsf{O}_e A$. Alternatively, one can interpret the logics in terms of our distinction between *prima facie* obligations and actual obligations (cf. Section 3.1).

Adaptive logics that are based on **P** itself are discussed in [43]; here we only discussed the second of the two. The other AL discussed by Goble appears to be slightly weaker. For instance, in this logic, the Natascha argument is only valid if we use *minimal abnormality.* More generally, in this logic any conflict of the type $\mathsf{O}A \wedge \mathsf{O}B \wedge \neg\mathsf{O}(A \wedge B)$ "infects" all the subformulas of A and B. We leave the full inspection and proof of this claim for another occasion.

An interesting issue concerns the enrichment of the aforementioned ALs with operators that allow one to express (technical, physical, practical) impossibility at the object level. Indeed, in Williams' famous essay, he argues that purely logical conflicts between oughts are only a special case of a much more common type of conflicts, viz. conflicts between two obligations whose joint fulfillment is impossible for *contingent* reasons – e.g. because of the particular physical situation we find ourselves in [111]. This raises a number of questions concerning the interplay between alethic and deontic modalities, which would take us well beyond the scope of the present paper – see however [15, Chapter 4] for a first attempt to combine alethic and deontic modalities.

7 Inconsistency-adaptive deontic logics

As noted in Section 3.5, the first adaptive logics were *inconsistency-adaptive.* These logics are members of the larger family of *paraconsistent logics*, i.e. logics which invalidate (ECQ).

Note that (ECQ) bears close affinity to (DEX). To obtain the latter from the former we only need to prefix the formulas involved with an O-operator. Besides the approaches we saw in Sections 5 and 6, a third natural way to invalidate (DEX) is by invalidating (ECQ).

Going paraconsistent has a couple of additional benefits in the context of deontic logic. A first is that it allows us to preserve the interdefinability of O and P, while invalidating (DEX-$\mathsf{O}\mathsf{P}\neg$). Assuming the interdefinability of O and P, the formula $\mathsf{O}A \wedge \mathsf{P}\neg A$ is equivalent to the contradictions $\mathsf{O}A \wedge \neg\mathsf{O}A$ and $\neg\mathsf{P}\neg A \wedge \mathsf{P}\neg A$. By (ECQ), these contradictions entail everything. To prevent such explosive behavior, it suffices to invalidate (ECQ).

A second advantage is that only a paraconsistent deontic logic can invalidate the explosion principles (DEX-$\mathsf{O}\neg\mathsf{O}$) and (DEX-$\mathsf{P}\neg\mathsf{P}$), for the obvious reason that these principles are instances of (ECQ):

$$\mathsf{O}A, \neg\mathsf{O}A \vdash \mathsf{O}B \tag{DEX-O¬O}$$

$$\mathsf{P}A, \neg\mathsf{P}A \vdash \mathsf{O}B \tag{DEX-P¬P}$$

There are independent reasons as to why, in some contexts, we may want to tolerate *contradictory norms*, i.e. formulas of the form $\mathsf{O}A \wedge \neg\mathsf{O}A$ or $\mathsf{P}A \wedge \neg\mathsf{P}A$. Priest, for instance, gives the following example. Suppose that, in some country, women are not permitted to vote, while property holders are permitted to vote. Suppose further that, perhaps due to a recent revision of the property law, women are permitted to hold property. Then female property holders are both permitted and not permitted to vote ($\mathsf{P}v \wedge \neg\mathsf{P}v$) [81, pp. 184–185].

In this section, we present inconsistency-adaptive deontic logics. We will work stepwise, starting with the paraconsistent logic **CLuN**, its deontic extension **DCLuN**, and adaptive strengthenings **DCLuN**$^{\mathbf{x}}$ (Section 7.1). After that, we will consider several variants of **DCLuN** and their associated adaptive logics (sections 7.2 and 7.3).

7.1 Paraconsistent adaptive deontic logic

A paraconsistent core logic We use the paraconsistent logic **CLuN** as our starting point. **CLuN** is an acronym for 'Classical Logic with gluts for Negation'. A truth-value *glut* for negation relative to a formula A occurs when both A and its negation are true; **CLuN** allows such gluts whereas **CL** disallows them. The deontic logics to be presented in this section are extensions of **CLuN**, but they are defined so that plenty of other paraconsistent logics may replace **CLuN** as their core logic. In Sections 7.2 and 7.3 we will mention some alternatives.

The set $\mathcal{W}^{\sim}$ of well-formed **CLuN**-formulas is the following:

$$\begin{aligned}\mathcal{W}^{\sim} := \;& \mathcal{S} \mid {\sim}\langle\mathcal{W}^{\sim}\rangle \mid \neg\langle\mathcal{W}^{\sim}\rangle \mid \langle\mathcal{W}^{\sim}\rangle \vee \langle\mathcal{W}^{\sim}\rangle \mid \langle\mathcal{W}^{\sim}\rangle \wedge \langle\mathcal{W}^{\sim}\rangle \mid \\ & \langle\mathcal{W}^{\sim}\rangle \supset \langle\mathcal{W}^{\sim}\rangle \mid \langle\mathcal{W}^{\sim}\rangle \equiv \langle\mathcal{W}^{\sim}\rangle\end{aligned}$$

In the remainder, we will stick to $\neg$ as the connective denoting classical negation. Beside $\neg$, $\mathcal{W}^{\sim}$ contains the connective $\sim$ which we will use as our paraconsistent negation sign. In fact, $\sim$ is the only **CLuN**-connective which behaves differently from the classical connectives. We obtain **CLuN** by adding the following axiom schema to **CL**:

$$A \vee {\sim}A \tag{EM$\sim$}$$

We write $\Gamma \vdash_{\mathbf{CLuN}} A$ to denote that A is **CLuN**-derivable from Γ.

The **CLuN**-semantics is defined as follows. To obtain a **CLuN**-model M, we extend the assignment function v_a of **CL** so that it assigns truth values not only to schematic letters, but also to formulas of the form $\sim A$, i.e. $v_a : \mathcal{S} \cup \{\sim A \mid A \in \mathcal{W}^{\sim}\} \to \{0, 1\}$. Next, we extend v_a to a valuation function v as follows:

(SC1) For formulas $A \in \mathcal{S} \cup \{\sim A \mid A \in \mathcal{W}^{\sim}\} : M \models A$ iff $v_a(A) = 1$.
(SC2) For $\neg, \vee, \wedge, \supset, \equiv$, the semantic clauses for **CLuN** are those of **CL**.

Finally, in order to validate the axiom (EM$\sim$), we require that all **CLuN**-models satisfy the following condition: for all $A \in \mathcal{W}^{\sim}$, $M \models A$ or $M \models \sim A$. A semantic consequence relation for **CLuN** is defined as follows: $\Gamma \Vdash_{\mathbf{CLuN}} A$ iff for all **CLuN**-models M: if $M \models B$ for all $B \in \Gamma$, then $M \models A$.

Before we move on to deontic extensions of **CLuN**, we point out a number of relevant properties of this logic for ease of reference:

(i) **CLuN** is paraconsistent, but not *paracomplete*: while (ECQ) is **CLuN**-invalid for $\sim$, the excluded middle principle (EM$\sim$) is **CLuN**-valid.

(ii) In contrast to well-known paraconsistent logics such as Priest's **LP**, **CLuN** validates modus ponens:

$$A, A \supset B \vdash B \tag{MP}$$

Note that $A \supset B$ and $\sim A \vee B$ are not **CLuN**-equivalent: if $v(A) = v(\sim A) = v(\sim B) = 1$ and $v(B) = 0$, then $v(A \supset B) = 0$ while $v(\sim A \vee B) = 1$.

(iii) De Morgan's laws and the double negation laws are invalid for $\sim$ in **CLuN**. This means that complex contradictions are not reducible to contradictions between elementary letters:

$$(p \wedge q) \wedge \sim(p \wedge q) \nvdash (p \wedge \sim p) \vee (q \wedge \sim q) \tag{36}$$

$$(p \vee q) \wedge \sim(p \vee q) \nvdash (p \wedge \sim p) \vee (q \wedge \sim q) \tag{37}$$

$$(p \supset q) \wedge \sim(p \supset q) \nvdash (p \wedge \sim p) \vee (q \wedge \sim q) \tag{38}$$

$$\sim\sim(p \wedge \sim p) \nvdash p \wedge \sim p \tag{39}$$

(iv) Contraposition, modus tollens, and disjunctive syllogism are invalid for $\sim$ in **CLuN**:

$$A \supset B \nvdash \sim B \supset \sim A \tag{40}$$

$$A \supset B, \sim B \nvdash \sim A \tag{41}$$

$$A \vee B, \sim A \nvdash B \tag{42}$$

A paraconsistent deontic logic A technically straightforward way to construct a deontic logic on the basis of **CLuN** is the following. First, we extend the language $\mathcal{W}^{\sim}$ with the deontic operator O, preventing nested occurrences of the deontic operator:

$$\begin{aligned}\mathcal{W}^{\sim}_{\mathsf{O}} := \;& \mathcal{W}^{\sim} \mid \mathsf{O}\langle\mathcal{W}^{\sim}\rangle \mid {\sim}\langle\mathcal{W}^{\sim}_{\mathsf{O}}\rangle \mid \neg\langle\mathcal{W}^{\sim}_{\mathsf{O}}\rangle \mid \langle\mathcal{W}^{\sim}_{\mathsf{O}}\rangle \vee \langle\mathcal{W}^{\sim}_{\mathsf{O}}\rangle \mid \langle\mathcal{W}^{\sim}_{\mathsf{O}}\rangle \wedge \\ & \langle\mathcal{W}^{\sim}_{\mathsf{O}}\rangle \mid \langle\mathcal{W}^{\sim}_{\mathsf{O}}\rangle \supset \langle\mathcal{W}^{\sim}_{\mathsf{O}}\rangle \mid \langle\mathcal{W}^{\sim}_{\mathsf{O}}\rangle \equiv \langle\mathcal{W}^{\sim}_{\mathsf{O}}\rangle\end{aligned}$$

The logic **DCLuN** is axiomatized by adding to **CLuN** the axioms (K), (D), and closing the resulting set under (N) and (MP). Note that for (D) we need the original version (cf. page 9), hence with classical negations ($\neg$) only.

The semantics for **DCLuN** looks as follows. A model is a quadruple $M = \langle W, w_0, R, v\rangle$ where W is a non-empty set, $w_0 \in W$, $R \subseteq W \times W$ is a serial accessibility relation, and $v : \mathcal{W}^{\sim}_{\mathsf{O}} \times W \to \{1, 0\}$ is a valuation function. As with **CLuN**, we first assign truth values to both schematic letters and formulas of the form ${\sim}A$: $v_a : \mathcal{S} \cup \{{\sim}A \mid A \in \mathcal{W}^{\sim}_{\mathsf{O}}\} \times W \to \{0, 1\}$. v_a is extended to v as follows:

(SC1') For formulas $A \in \mathcal{S} \cup \{{\sim}A \mid A \in \mathcal{W}^{\sim}_{\mathsf{O}}\}$: $M, w \models A$ iff $v_a(A, w) = 1$.
(SC2') For $\mathsf{O}, \neg, \vee, \wedge, \supset, \equiv$, the semantic clauses for **DCLuN** are exactly those of **SDL** (cf. Section 2).

A model M is a **DCLuN**-*model* iff it satisfies the following condition on v:

$$\text{for all } w \in W, \text{ for all } A : v(A, w) = 1 \text{ or } v({\sim}A, w) = 1 \qquad (\mathrm{C}_u)$$

$\Gamma \Vdash_{\mathbf{DCLuN}} A$ iff for all **DCLuN**-models M: if $M, w_0 \models B$ for all $B \in \Gamma$, then $M, w_0 \models A$.

The proof of soundness for this logic is a matter of routine. For completeness, we can use the well-known technique of canonical models (see e.g. [24, Chapter 4]), adjusted to the setting with an actual world. Fix a maximal, $\neg$-consistent set $\Gamma \subseteq \mathcal{W}^{\sim}_{\mathsf{O}}$. We build the canonical model $M^c_\Gamma = \langle W^c, \Gamma, R^c, V^c\rangle$ for this set as follows:

(i) W^c is the set of all maximal consistent and **DCLuN**-closed sets Δ,

(ii) $R^c = \{(\Delta, \Delta') \mid \{A \mid \mathsf{O}A \in \Delta\} \subseteq \Delta'\}$,

(iii) for all $A \in \mathcal{S} \cup \{{\sim}A \mid A \in \mathcal{W}^{\sim}_{\mathsf{O}}\}$, for all $\Delta \in W^c$: $v_a(A, \Delta) = 1$ iff $A \in \Delta$.

To show that M^c_Γ is a **DCLuN**-model, we need to rely on excluded middle for $\sim$ and the maximality of each $\Delta \in W^c$. For seriality, we rely on the (D)-axiom in the usual way. The proof of the truth lemma proceeds by a standard induction. So we can derive that all the members of Γ are satisfied at Γ in M^c_Γ.

Note that, since **CLuN** is a conservative extension of **CL**, **DCLuN** is also a conservative extension of **SDL**. However, if we consider the $\neg$-free fragment of

DCLuN, and treat $\sim$ as the "proper" negation, then **DCLuN** is a proper fragment of **SDL**. When applying the logic **DCLuN** to concrete examples, we will use $\sim$ to translate negations in natural language. Given this convention, the logic **DCLuN** is strongly conflict-tolerant.

$$\mathsf{O}A \wedge \mathsf{O}{\sim}A \nvdash_{\mathbf{DCLuN}} \mathsf{O}B$$
$$\mathsf{O}A \wedge {\sim}\mathsf{O}A \nvdash_{\mathbf{DCLuN}} \mathsf{O}B$$

In **DCLuN** we can define permission in various ways relative to our negation operators:

$$\mathsf{P}^{\neg}_{\neg}A =_{\mathsf{df}} \neg\mathsf{O}\neg A$$
$$\mathsf{P}^{\neg}_{\sim}A =_{\mathsf{df}} \neg\mathsf{O}{\sim}A$$
$$\mathsf{P}^{\sim}_{\neg}A =_{\mathsf{df}} {\sim}\mathsf{O}\neg A$$
$$\mathsf{P}^{\sim}_{\sim}A =_{\mathsf{df}} {\sim}\mathsf{O}{\sim}A$$

All of these permission operators tolerate conflicts between an obligation and a permission, as well as contradictory norms. Where $\dagger, \ddagger \in \{\sim, \neg\}$:

$$\mathsf{O}A \wedge \mathsf{P}^{\ddagger}_{\dagger}{\sim}A \nvdash_{\mathbf{DCLuN}} \mathsf{O}B \tag{43}$$
$$\mathsf{O}{\sim}A \wedge \mathsf{P}^{\ddagger}_{\dagger}A \nvdash_{\mathbf{DCLuN}} \mathsf{O}B \tag{44}$$
$$\mathsf{P}^{\ddagger}_{\dagger}A \wedge {\sim}\mathsf{P}^{\ddagger}_{\dagger}A \nvdash_{\mathbf{DCLuN}} \mathsf{O}B \tag{45}$$

In sum, **DCLuN** is very conflict-tolerant, especially compared to the logics discussed in previous sections. However, it is also rather weak. To be sure, the Jones argument, the Roberts arguments, and the (original and modified) Natascha argument are valid in **DCLuN** due to the validity of (Inh) and (Agg). Unfortunately, the Smith argument and the Thomas argument are not **DCLuN**-valid. More generally, all instances of the following inference schemas fail in **DCLuN**:

$$\mathsf{O}(A \supset B) \nvdash_{\mathbf{DCLuN}} \mathsf{O}({\sim}B \supset {\sim}A) \tag{46}$$
$$\mathsf{O}(A \supset B), \mathsf{O}{\sim}B \nvdash_{\mathbf{DCLuN}} \mathsf{O}{\sim}A \tag{47}$$
$$\mathsf{O}(A \vee B), \mathsf{O}{\sim}A \nvdash_{\mathbf{DCLuN}} \mathsf{O}B \tag{48}$$

The invalidity of (46)-(48) mirrors the invalidity of their non-deontic counterparts (40)-(42) in **CLuN**. So the main advantage of **DCLuN** goes hand in hand with its inability to validate seemingly intuitive inferences. This drawback is overcome by strengthening this system within the adaptive logics framework.

Going adaptive We strengthen **DCLuN** to the adaptive logic **DCLuN**$^{\mathbf{x}}$, which is defined by the triple $\langle \mathbf{DCLuN}, \Omega^{\sim}, \mathsf{x}\rangle$, where

$$\Omega^{\sim} = \{A \wedge {\sim}A \mid A \in \mathcal{W}^{\sim}_{\mathsf{O}}\} \cup \{\mathsf{P}^{\neg}_{\neg}(A \wedge {\sim}A) \mid A \in \mathcal{W}^{\sim}\}$$

$\Omega^{\sim}$ contains not only plain contradictions, but also formulas that express that in some deontically accessible world, a given contradiction is true. This allows us at once to validate the Smith argument and the Thomas argument. Here is a **DCLuN**$^{\mathbf{x}}$-proof illustrating the validity of the Thomas argument:

1	$\mathsf{O}(t \wedge (f \vee s))$	Prem	$\emptyset$
2	$\mathsf{O}({\sim}t \wedge {\sim}f)$	Prem	$\emptyset$
3	$\mathsf{O}(f \vee s)$	1; RU	$\emptyset$
4	$\mathsf{O}{\sim}f$	2; RU	$\emptyset$
5	$\mathsf{O}s$	3,4; RC	$\{\mathsf{P}^{\neg}_{\neg}(f \wedge {\sim}f)\}$

The inference made at line 5 holds in view of the **DCLuN**-valid inference

$$\mathsf{O}(f \vee s), \mathsf{O}{\sim}f \vdash \mathsf{O}s \vee \mathsf{P}^{\neg}_{\neg}(f \wedge {\sim}f) \tag{49}$$

Suppose that $\mathsf{O}(f \vee s)$ and $\mathsf{O}{\sim}f$. By (Agg), $\mathsf{O}((f \vee s) \wedge {\sim}f)$. By normal modal logic properties, we can infer $\mathsf{O}s \vee \neg\mathsf{O}\neg(f \wedge {\sim}f)$ so that we can derive $\mathsf{O}s$ on the condition $\mathsf{P}^{\neg}_{\neg}(f \wedge {\sim}f)$.

Equations (50)-(55) illustrate that the **DCLuN**-invalid inferences (40)-(42) and (46)-(48) hold conditionally in **DCLuN**$^{\mathbf{x}}$. The conditions on which these inferences can be made in a **DCLuN**$^{\mathbf{x}}$-proof are indicated between square brackets.

$$p \supset q \vdash_{\mathbf{DCLuN}^{\mathbf{x}}} {\sim}q \supset {\sim}p \qquad [q \wedge {\sim}q] \tag{50}$$

$$p \supset q, {\sim}q \vdash_{\mathbf{DCLuN}^{\mathbf{x}}} {\sim}p \qquad [q \wedge {\sim}q] \tag{51}$$

$$p \vee q, {\sim}p \vdash_{\mathbf{DCLuN}^{\mathbf{x}}} q \qquad [p \wedge {\sim}p] \tag{52}$$

$$\mathsf{O}(p \supset q) \vdash_{\mathbf{DCLuN}^{\mathbf{x}}} \mathsf{O}({\sim}q \supset {\sim}p) \qquad [\mathsf{P}^{\neg}_{\neg}(q \wedge {\sim}q)] \tag{53}$$

$$\mathsf{O}(p \supset q), \mathsf{O}{\sim}q \vdash_{\mathbf{DCLuN}^{\mathbf{x}}} \mathsf{O}{\sim}p \qquad [\mathsf{P}^{\neg}_{\neg}(q \wedge {\sim}q)] \tag{54}$$

$$\mathsf{O}(p \vee q), \mathsf{O}{\sim}p \vdash_{\mathbf{DCLuN}^{\mathbf{x}}} \mathsf{O}q \qquad [\mathsf{P}^{\neg}_{\neg}(p \wedge {\sim}p)] \tag{55}$$

More generally, relative to premise sets from which no abnormalities are **DCLuN**-derivable $\sim$ is as strong as $\neg$ in **DCLuN**$^{\mathbf{x}}$. That is, where $A \in \mathcal{W}^{\sim}_{\mathsf{O}}$, let $\pi(A)$ be the result of replacing every occurrence of $\sim$ in A with $\neg$. We lift this translation to sets of formulas in the usual way. We can now prove the following:

Theorem 7.1. *If* Γ *is normal, then* $\Gamma \vdash_{\mathbf{DCLuN}^{\mathbf{x}}} A$ *iff* $\pi(\Gamma) \vdash_{\mathbf{SDL}} \pi(A)$.

Proof. The upper limit logic of $\mathbf{DCLuN^x}$ is obtained by adding to **DCLuN** all formulas $\neg A$ for which $A \in \Omega^{\sim}$. Call this logic **UDCLuN**. By Theorem 3.14: If Γ is normal, then $\Gamma \vdash_{\mathbf{DCLuN^x}} A$ iff $\Gamma \vdash_{\mathbf{UDCLuN}} A$. We show that $\Gamma \vdash_{\mathbf{UDCLuN}} A$ iff $\pi(\Gamma) \vdash_{\mathbf{SDL}} \pi(A)$.
($\Rightarrow$) It is easily checked that, under the transformation given, all **CLuN**-valid inferences are **CL**-valid; (K), (D), and (N) are **SDL**-valid; and all elements of $\pi(\{\neg A \mid A \in \Omega^{\sim}\})$ are **SDL**-valid.
($\Leftarrow$) Given the fact that **UDCLuN**, like **DCLuN**, extends **SDL**, it suffices to show that $\sim$ is as strong as $\neg$ in **UDCLuN**:

$$\vdash_{\mathbf{UDCLuN}} {\sim}A \supset \neg A \tag{56}$$

$$\vdash_{\mathbf{UDCLuN}} \mathsf{O}{\sim}A \supset \mathsf{O}\neg A \tag{57}$$

Ad. (56) Suppose ${\sim}A$. Then $\neg A \vee (A \wedge {\sim}A)$ since $\vdash_{\mathbf{CLuN}} {\sim}A \supset (\neg A \vee (A \wedge {\sim}A))$. We also know that $\vdash_{\mathbf{UDCLuN}} \neg(A \wedge {\sim}A)$, so by **CL**-properties we obtain $\neg A$.
*Ad. (57)*By (N), $\vdash_{\mathbf{UDCLuN}} \mathsf{O}({\sim}A \supset (\neg A \vee (A \wedge {\sim}A)))$. Suppose $\mathsf{O}{\sim}A$. By (K) and (MP), $\mathsf{O}(\neg A \vee (A \wedge {\sim}A))$. By **SDL**-properties, $\mathsf{O}\neg A \vee \mathsf{P}^{\neg}_{\neg}(A \wedge {\sim}A)$. But then $\mathsf{O}\neg A$ follows in view of $\vdash_{\mathbf{UDCLuN}} \neg\mathsf{P}^{\neg}_{\neg}(A \wedge {\sim}A)$. □

7.2 Semi-paraconsistent adaptive deontic logic

The logic **DCLuN** and its adaptive extensions consistently accommodate all types of normative conflicts that we have encountered so far. But they also consistently accommodate plain contradictions between formulas not involving deontic operators, such as $p \wedge {\sim}p$. One could argue that this is overkill. Even if normative conflicts are part of life and should be accommodated in a deontic logic, there is no need to allow also for a non-deontic statement and its negation to be true at the same time.

In this section we mention two ways to adjust **DCLuN** and its adaptive extensions so as to tolerate normative conflicts, without having to tolerate all outright contradictions of the form $A \wedge {\sim}A$. Casey McGinnis coined the term *semi-paraconsistent deontic logic* for paraconsistent deontic logics that meet this desideratum [64; 63].

Excluding non-deontic contradictions The logic $\mathbf{DCLuN_1}$ is obtained by closing **DCLuN** under the axiom schema (Cons_1):[53]

$$\text{Where } A \in \mathcal{W}^{\sim} : {\sim}A \supset \neg A \tag{Cons_1}$$

[53] Where $\vdash\ \subseteq \wp(\Phi) \times \Phi$ is a consequence relation and Δ is a set of axioms, we obtain $\vdash_{\Delta}$, *the closure of* $\vdash$ *under* Δ, as follows: $\Gamma \vdash_{\Delta} A$ iff $\Gamma \cup \Delta \vdash A$. This means that one cannot e.g. apply necessitation to members of Δ.

Where $A \in \mathcal{W}^{\sim}$, (Cons$_1$) takes care that $A \wedge {\sim}A$ is trivialized in $\mathbf{DCLuN_1}$. This means that for non-deontic formulas, we obtain full **CL**. Still, $\mathbf{DCLuN_1}$, like **DCLuN**, is highly conflict-tolerant. Where as before $\dagger, \ddagger \in \{\sim, \neg\}$:

$$\mathsf{O}A \wedge \mathsf{O}{\sim}A \nvdash_{\mathbf{DCLuN_1}} \mathsf{O}B \tag{58}$$

$$\mathsf{O}A \wedge \mathsf{P}^{\ddagger}_{\dagger}{\sim}A \nvdash_{\mathbf{DCLuN_1}} \mathsf{O}B \tag{59}$$

$$\mathsf{O}{\sim}A \wedge \mathsf{P}^{\ddagger}_{\dagger}A \nvdash_{\mathbf{DCLuN_1}} \mathsf{O}B \tag{60}$$

$$\mathsf{O}A \wedge {\sim}\mathsf{O}A \nvdash_{\mathbf{DCLuN_1}} \mathsf{O}B \tag{61}$$

$$\mathsf{P}^{\ddagger}_{\dagger}A \wedge {\sim}\mathsf{P}^{\ddagger}_{\dagger}A \nvdash_{\mathbf{DCLuN_1}} \mathsf{O}B \tag{62}$$

As desired, $\mathbf{DCLuN_1}$ consistently accommodates normative conflicts while trivializing contradictions between statements without occurrences of deontic operators.

Semantically, the logic $\mathbf{DCLuN_1}$ is characterized by imposing the following additional condition on **DCLuN**-models:

$$\text{For all } A \in \mathcal{W}^{\sim} : v(A, w_0) = 1 \text{ iff } v({\sim}A, w_0) = 0 \tag{C$^0_1$}$$

Unlike **DCLuN**, the logic $\mathbf{DCLuN_1}$ is not a normal modal logic, since it is not closed under the standard necessitation rule (N). That is, even though ${\sim}p \supset \neg p$ is a theorem of the logic, $\mathsf{O}({\sim}p \supset \neg p)$ is not. For similar reasons, the logic is not closed under Uniform Substitution. For instance, ${\sim}\mathsf{O}p \supset \neg\mathsf{O}p$ is not a theorem of $\mathbf{DCLuN_1}$.

Adaptive logics based on $\mathbf{DCLuN_1}$ can be defined just as before. Mind however that abnormalities of the form $A \wedge {\sim}A$ for $A \in \mathcal{W}^{\sim}$ are vacuous in the resulting adaptive logics, since they are anyway trivialized by their lower limit logic, in view of (Cons$_1$). These adaptive logics will perform just as well as $\mathbf{DCLuN^x}$, in that they validate all the inferences from our list of benchmark examples.

Excluding all contradictions at the actual world A second, stronger semi-paraconsistent deontic logic is obtained by closing **DCLuN** under the unrestricted version of (Cons$_1$):

$${\sim}A \supset \neg A \tag{Cons$_2$}$$

Call the resulting logic $\mathbf{DCLuN_2}$. Its semantics is obtained by imposing the following condition on **DCLuN**-models:

$$v({\sim}A, w_0) = 1 \text{ iff } v(A, w_0) = 0 \tag{C$^0_2$}$$

In the $\mathbf{DCLuN_2}$-semantics, $\sim$ and $\neg$ are interchangeable at w_0. At all other worlds, $\neg$ remains strictly stronger than $\sim$. This means that contradictions outside the scope of O are trivialized, whereas contradictions within the scope of O are not.

The logic $\mathbf{DCLuN_2}$ is not as conflict-tolerant as $\mathbf{DCLuN_1}$, since it trivializes conflicts of the form $\mathsf{O}A \wedge {\sim}\mathsf{O}A$ or $\mathsf{P}^{\ddagger}_{\dagger}A \wedge {\sim}\mathsf{P}^{\ddagger}_{\dagger}A$, where $\dagger, \ddagger \in \{\sim, \neg\}$. Since (Cons$_2$) and (C$^0_2$) are no longer restricted to members of $\mathcal{W}^{\sim}$, the logic $\mathbf{DCLuN_2}$ satisfies the rule of uniform substitution, although necessitation (in its full generality) is still invalid.

Just as with **DCLuN** and $\mathbf{DCLuN_1}$, we can use $\mathbf{DCLuN_2}$ as a lower limit logic of our adaptive logic. In this case, the set of abnormalities can be further simplified to the following:

$$\Omega^{\sim}_2 = \{\mathsf{P}^{\sim}_{\neg}(A \wedge {\sim}A) \mid A \in \mathcal{W}^{\sim}\}$$

7.3 Other paraconsistent negations

CLuN is the weakest logic which verifies the full positive fragment of **CL** as well as the principle of Excluded Middle (EM). Stronger paraconsistent logics can be obtained by adding to **CLuN** the double negation laws and/or de Morgan's laws for negation:

$$\begin{aligned}
{\sim}{\sim}A &\equiv A && (\text{A}{\sim}{\sim})\\
{\sim}(A \supset B) &\equiv (A \wedge {\sim}B) && (\text{A}{\sim}{\supset})\\
{\sim}(A \wedge B) &\equiv ({\sim}A \vee {\sim}B) && (\text{A}{\sim}{\wedge})\\
{\sim}(A \vee B) &\equiv ({\sim}A \wedge {\sim}B) && (\text{A}{\sim}{\vee})\\
{\sim}(A \equiv B) &\equiv ((A \vee B) \wedge ({\sim}A \vee {\sim}B)) && (\text{A}{\sim}{\equiv})
\end{aligned}$$

Let **CLuNs** be obtained by adding all of these axioms to **CLuN**. Analogously to the construction of **DCLuN**, we can now construct the logic **DCLuNs** by enriching **CLuNs** with (K), (D), and (N).

One clear difference between **DCLuN**-based ALs and **DCLuNs**-based ALs is that the latter verify a number of additional inferences in a non-defeasible way. For instance, where $\Gamma = \{\mathsf{O}(p \wedge q), \mathsf{O}{\sim}(p \wedge q)\}$, one cannot $\mathbf{DCLuN^r}$-derive $\mathsf{O}({\sim}p \vee {\sim}q)$ from Γ, since one cannot rely on the falsehood of the abnormality $\mathsf{P}^{\sim}_{\neg}((p \wedge q) \wedge {\sim}(p \wedge q))$. In contrast, one can finally $\mathbf{DCLuNs^r}$-derive $\mathsf{O}({\sim}p \vee {\sim}q)$ from the same premise set, simply in view of properties of **DCLuNs**.

We have to take care when constructing adaptive logics on the basis of **DCLuNs**. Suppose that we work with the set $\Omega^{\sim}$ of $\mathbf{DCLuN^x}$-abnormalities.

1	$\mathsf{O}p$	Prem	$\emptyset$
2	$\mathsf{O}{\sim}p$	Prem	$\emptyset$
3	$\mathsf{O}q$	Prem	$\emptyset$
4	$\mathsf{O}({\sim}q \vee r)$	Prem	$\emptyset$
5	$\mathsf{O}r$	3,4;RC	$\{\mathsf{P}^{\urcorner}_{\neg}(q \wedge {\sim}q)\}\checkmark^{6}$
6	$\mathsf{P}^{\urcorner}_{\neg}(q \wedge {\sim}q) \vee \mathsf{P}^{\urcorner}_{\neg}((p \wedge r) \wedge {\sim}(p \wedge r))$	1-4;RU	$\emptyset$

Line 5 is marked in view of the minimal Dab-formula derived at line 6. There is no extension of this proof in which to unmark line 5. The proof illustrates that $\mathsf{O}r$ is not finally derivable from the premises at lines 1-4. This is counter-intuitive.

If we are to build an adaptive logic on the basis of the lower limit logic **DCLuNs** and the set of abnormalities $\Omega^{\sim}$, the resulting logic would exhibit flip-flop behavior (see Section 5 where we also encountered this problem). The solution is to restrict the set of abnormalities as follows:

$$\Omega^{\sim}_{s} = \{A \wedge {\sim}A \mid A \in \mathcal{S}\} \cup \{\mathsf{O}A \wedge {\sim}\mathsf{O}A \mid A \in \mathcal{W}^{\sim}\} \cup \{\mathsf{P}^{\urcorner}_{\neg}(A \wedge {\sim}A) \mid A \in \mathcal{S}\} \qquad (63)$$

Given (A$\sim\sim$)-(A$\sim$ $\equiv$), inconsistencies between complex formulas in $\mathcal{W}$ can be reduced to inconsistencies at the level of atoms in **DCLuNs**. In view of this, **DCLuNs$^{\mathsf{x}}$**-abnormalities must be restricted accordingly, on pain of flip-flop behavior. That is, where $A \in \mathcal{W}$, $A \wedge {\sim}A$ and $\mathsf{P}^{\urcorner}_{\neg}(A \wedge {\sim}A)$ only counts as an abnormality when $A \in \mathcal{S}$.

The situation is different for formulas of the form $\mathsf{O}A \wedge {\sim}\mathsf{O}A$: within the scope of O, inconsistencies between complex formulas do *not* reduce to inconsistencies at the level of atoms. For instance, the inference from $\mathsf{O}(p \wedge q) \wedge {\sim}\mathsf{O}(p \wedge q)$ to $(\mathsf{O}p \wedge {\sim}\mathsf{O}p) \vee (\mathsf{O}q \wedge {\sim}\mathsf{O}q)$ is not **DCLuNs**-valid, since ${\sim}\mathsf{O}(p \wedge q)$ does not **DCLuNs**-entail ${\sim}\mathsf{O}p \vee {\sim}\mathsf{O}q$. More generally, where A is a complex formula, the formula ${\sim}\mathsf{O}A$ cannot be further analysed in **DCLuNs**. So, as in **DCLuN$^{\mathsf{x}}$**, all formulas of the form $\mathsf{O}A \wedge {\sim}\mathsf{O}A$ count as abnormalities in **DCLuNs$^{\mathsf{x}}$**.

Let **DCLuNs$^{\mathsf{x}}$** be the adaptive logic defined by the lower limit logic **DCLuNs**, the set of abnormalities $\Omega^{\sim}_{s}$, and the strategy $\mathsf{x} \in \{\mathsf{r}, \mathsf{m}\}$. Then clearly the formula derived at line 6 of the proof above is no longer a minimal Dab-formula, and line 5 remains unmarked. We can still derive the Dab-formula $\mathsf{P}^{\urcorner}_{\neg}(q \wedge {\sim}q) \vee \mathsf{P}^{\urcorner}_{\neg}(p \wedge {\sim}p) \vee \mathsf{P}^{\urcorner}_{\neg}(r \wedge {\sim}r)$ from lines 1-4 via RU, in view of

$$\mathsf{P}^{\urcorner}_{\neg}((p \wedge r) \wedge {\sim}(p \wedge r)) \vdash_{\mathbf{DCLuNs}} \mathsf{P}^{\urcorner}_{\neg}(p \wedge {\sim}p) \vee \mathsf{P}^{\urcorner}_{\neg}(r \wedge {\sim}r) \qquad (64)$$

However, this Dab-formula is not minimal, since its disjunct $\mathsf{P}^{\urcorner}_{\neg}(p \wedge {\sim}p)$ is a **DCLuNs**-consequence of the formulas $\mathsf{O}p$ and $\mathsf{O}{\sim}p$ at lines 1 and 2. As a result, line 5 is finally derivable and $\mathsf{O}r$ is a **DCLuNs$^{\mathsf{x}}$**-consequence of the premises.

Other than **CLuN** and **CLuNs**, there is a wide variety of paraconsistent logics that can serve as the core logic of an inconsistency-adaptive logic. We could, for instance, treat '$\sim$' as a dummy operator for which not even (EM) holds by removing (A$\sim$1) in the axiomatization of **CLuN**. The resulting logic is called **CLoN** (for Classical Logic with both gluts and gaps for Negation). Extending **CLoN** with (A$\sim\sim$)-(A$\sim\equiv$) results in the logic **CLoNs**. These systems too can be extended deontically and adaptively. In addition, one can also consider semi-paraconsistent versions of **DCLuNs** and **DCLoNs**.

7.4 Further reading and open ends

For a general overview of paraconsistent logic, see e.g. [79; 80]. For an overview of (monotonic) paraconsistent deontic logic, we refer to [42, Sect. 6.1].

The first paper on inconsistency-adaptive logic – published in 1989, but written in 1981 – is [5], where the proof theory for the reliability strategy was first presented. The minimal abnormality strategy was first presented (semantically) in [4]. The (propositional) results of the two aforementioned papers were generalized to the predicative level in [7]. For an overview and more recent results within the inconsistency-adaptive program, see [12].

Inconsistency-adaptive deontic logics were presented in [15; 22], in [21], and in [43]. Most of these systems – in contrast to the ones presented in this section – allow for the following inference:[54]

$$\mathsf{O}A \wedge \mathsf{O}{\sim}A \vdash {\sim}\mathsf{O}{\sim}A \wedge \mathsf{O}{\sim}A \tag{65}$$

That is, conflicts of the form $\mathsf{O}A \wedge \mathsf{O}{\sim}A$ entail plain contradictions. Goble is critical of such systems:

> That seems an exceedingly strong commitment. It is easy to accept that there are normative conflicts, harder to suppose they all yield contradictions that are true. Even Priest, the hierarch of dialetheism, does not consider normative conflicts so paradoxical [43, Fn. 15].

The systems presented in this section circumvent Goble's criticism by invalidating inferences like (65).

In [16] the semi-paraconsistent deontic logic **LNP** is presented and extended within the adaptive logics framework. **LNP** is a close cousin of $\mathbf{DCLoNs_2}$, but has

[54] (65) holds for the inconsistency-adaptive systems presented in [15; 22], and [21]. The closely related principle $\mathsf{O}A \wedge \mathsf{O}{\sim}A \vdash (\mathsf{O}{\sim}A \wedge {\sim}\mathsf{O}{\sim}A) \vee \mathsf{O}B$ holds for those logics mentioned in [43] which satisfy the 'deontic addition' schema $\mathsf{O}A \supset \mathsf{O}(A \vee B)$.

a slightly different language in which the P-operator is primitive, and in which '$\neg$' is allowed only outside the scope of deontic operators, while '$\sim$' is allowed only inside the scope of deontic operators.

Once we are open to the possibility of changing the logic of the connectives, new questions arise. For instance, why should we always blame negation for the explosive behavior of a logic, and why not weaken the meaning of the other connectives? Why not e.g. give up addition for $\vee$ (i.e., to derive $A \vee B$ from A or from B)? In [8], Batens shows that a whole range of interesting new logics come to the fore, once we generalize the idea of gluts and gaps to other connectives and logical operators. The application of all this to deontic reasoning is yet to be studied in detail, but it can draw on many existing results concerning corrective ALs.

In [17], a very rich paraconsistent deontic logic is presented, one that allows the user to express not only obligations that concern states of affairs, but also obligations that concern agency. The language of these systems contains modal operators $\Box_J$ for "the group of agents J brings it about that", inspired by existing work on logics of agency [86; 23; 31]. This in turn allows one to distinguish between various different types of *inter-personal* and *intra-personal* deontic conflicts:[55]

$$\mathsf{O}\Box_i A \wedge \mathsf{O}\Box_j {\sim} A \tag{66}$$

$$\mathsf{O}\Box_i A \wedge \mathsf{P}\Box_j {\sim} A \tag{67}$$

$$\mathsf{O}\Box_i A \wedge \mathsf{O}\Box_i {\sim} A \tag{68}$$

$$\mathsf{O}\Box_i A \wedge \mathsf{P}\Box_i {\sim} A \tag{69}$$

$$\mathsf{O}\Box_i A \wedge \mathsf{O}{\sim}\Box_i A \tag{70}$$

$$\mathsf{O}\Box_i A \wedge \mathsf{P}{\sim}\Box_i A \tag{71}$$

$$\mathsf{O}\Box_i A \wedge {\sim}\mathsf{O}\Box_i A \tag{72}$$

$$\mathsf{P}\Box_i A \wedge {\sim}\mathsf{P}\Box_i A \tag{73}$$

One further advantage of such richer formal languages in the context of adaptive reasoning is that they allow us to prioritize the minimization of certain types of conflicts over that of others. For instance, we may consider conflicts of type (68) worse than those of type (66) and (67), since the former clearly violate the principle that if an agent ought to bring about A, then that agent is also able to see to A

[55] An inter-personal conflict is one that holds between the obligations of different agents, whereas an intra-personal conflict obtains between the obligations of a single agent. One famous example of an inter-personal normative conflict can be found in Sophocles' *Antigone*, where due to the city's laws, Creon is obliged to prevent the burial or Antigon's brother Polyneices, but Antigone faces a religious and familial obligation to bury Polyneices [62; 44].

– assuming agents cannot bring about contradictions. Such a prioritized reasoning can be modeled in terms of a lexicographic AL (cf. Section 3.4).

8 Conflict-tolerant adaptive logics: round-up

In this section, we give an overview of the main features of the logics discussed so far. We start by giving an overview of the performance of revisionist ALs with respect to the criteria introduced in Section 4.2. In Section 8.2 we return to the logics from Section 3. We show how these can be evaluated using similar criteria, and how they can be enriched in various ways.

8.1 Revisionist deontic adaptive logics: overview

The behavior of the revisionist adaptive logics with respect to the criteria from Section 4.2 is summarized in Tables 1 and 2. Principles (arguments) that are valid in a given logic receive a 3, invalid principles (arguments) receive a 7.[56] Where the premises of an argument are trivialized by a given logic, we write a $\perp$ in Table 2.

	DEX	DEX-$\mathsf{O}\perp$	DEX-$\mathsf{P}\perp$	DEX-$\mathsf{OP}\neg$	DEX-$\mathsf{O}\neg\mathsf{P}$
$\mathbf{LUM.a}^{\mathbf{x}}$	✗	✗	✓	✓	✗
$\mathbf{LUM.b}^{\mathbf{x}}$	✗	✓	✓	✓	✗
$\mathbf{LUM.c}^{\mathbf{x}}$	✗	✓	✓	✓	✗
$\mathbf{P}^{\mathbf{x}}$	✗	✓	✓	✓	✗
$\mathbf{DCLuN}^{\mathbf{x}}$	✗	✗	✗	✗	✗
$\mathbf{DCLuN}^{\mathbf{x}}_{\mathbf{1}}$	✗	✗	✗	✗	✗
$\mathbf{DCLuN}^{\mathbf{x}}_{\mathbf{2}}$	✗	✗	✗	✗	✓

Table 1: Behavior of deontic ALs with respect to various explosion principles.

It should be noted here once more (in line with our remarks in Section 4.2) that whether a given AL validates some form of deontic explosion or a specific inference should not be seen as conclusive evidence in favour of or against such a logic. The above tables are mostly for purposes of comparison and classification, and do not serve as strict criteria of the relative success or failure of the respective systems or their purposes. For example, with a view to supporting ought-implies-can, a system might be *designed* to consider $\mathsf{O}(A \wedge \neg A)$ inconsistent even while $\mathsf{O}A \wedge \mathsf{O}\neg A$ is

[56] As noted before, for the logics from Section 7 we assume that the principles (arguments) in question are formalized using the paraconsistent negation sign $\sim$.

	S	J	R1	R2	T	N1	N2
$\mathbf{LUM.a^{r}}$	✓	✓	✗	✗	✗	⊥	✗
$\mathbf{LUM.a^{m}}$	✓	✓	✗	✗	✗	⊥	✓
$\mathbf{LUM.b^{r}}$	✓	✓	✗	✗	✗	⊥	✗
$\mathbf{LUM.b^{m}}$	✓	✓	✗	✗	✗	⊥	✓
$\mathbf{LUM.c^{r}}$	✓	✓	✗	✗	✗	✗	✗
$\mathbf{LUM.c^{m}}$	✓	✓	✗	✗	✗	✓	✓
$\mathbf{P^{x}}$	✓	✓	✓	✓	✓	✓	✓
$\mathbf{DCLuN^{x}}$	✓	✓	✓	✓	✓	✓	✓
$\mathbf{DCLuN^{x}_{1}}$	✓	✓	✓	✓	✓	✓	✓
$\mathbf{DCLuN^{x}_{2}}$	✓	✓	✓	✓	✓	⊥	✓
$\mathbf{DCLuNs^{x}}$	✓	✓	✓	✓	✓	✓	✓

Table 2: Behavior of deontic ALs with respect to the Smith (S), Jones (J), Roberts (R1 and R2), Thomas (T), and Natascha (N1 and N2) arguments from Section 4.)

consistent. In that case, that the system validates (DEX-$\mathsf{O}\bot$) may be taken as a virtue rather than a vice. Likewise, the validation of (DEX-$\mathsf{OP}\neg$) would be embraced by one with a classical point of view (and given the standard interdefinability of O and P).

Let us close this overview with a technical point. All ALs discussed in Sections 5–7 have a monotonic, conflict-tolerant deontic logic as their lower limit logic. The latter logics are mutually incomparable, in the sense that none is stronger than any other.[57] For instance, the logic **LUM.a** from Section 5 invalidates (Inh) but validates (Agg); conversely, the logic **P** that is discussed in Section 6 invalidates (Agg) but validates (Inh). It can easily be shown that any two ALs that are based on such incomparable lower limit logics, are themselves equally incomparable. This is an immediate corollary of the following:[58]

Theorem 8.1. *Let* $\mathbf{AL_1}$ *and* $\mathbf{AL_2}$ *be two ALs in standard format, defined by the triples* $\langle \mathbf{LLL_1}, \Omega_1, \mathsf{x}_1\rangle$*, resp.* $\langle \mathbf{LLL_2}, \Omega_2, \mathsf{x}_2\rangle$*, over a given formal language. If* $\vdash_{\mathbf{AL_1}} \subseteq \vdash_{\mathbf{AL_2}}$*, then* $\vdash_{\mathbf{LLL_1}} \subseteq \vdash_{\mathbf{LLL_2}}$*.*

Proof. By contraposition: suppose that $\vdash_{\mathbf{LLL_1}} \not\subseteq \vdash_{\mathbf{LLL_2}}$. Let Γ, A be such that

[57] A small warning is in place here. The paraconsistent deontic logics of the **DCLuN**-family, presented in Section 7, work with a richer language that contains both a paraconsistent and a classical negation. The claim we make here concerns the fragment of those logics *without* the classical negation.

[58] Theorem 8.1 generalizes one direction of Theorem 3.3 in [104].

(i) $\Gamma \vdash_{\mathbf{LLL_1}} A$ but (ii) $\Gamma \nvdash_{\mathbf{LLL_2}} A$. By (i) and the monotonicity of $\mathbf{LLL_1}$, (iii) $\Gamma \cup \{\neg A\} \vdash_{\mathbf{LLL_1}} A$. by (ii), $\Gamma \cup \{\neg A\}$ is $\mathbf{LLL_2}$-consistent, and hence by **CL**-properties, (iv) $\Gamma \cup \{\neg A\} \nvdash_{\mathbf{LLL_2}} A$. By (iii) and Theorem 3.15, $\Gamma \cup \{\neg A\} \vdash_{\mathbf{AL_1}} A$. By (iv) and Theorem 3.11, $\Gamma \cup \{\neg A\} \nvdash_{\mathbf{AL_2}} A$. Hence, $\vdash_{\mathbf{AL_1}} \not\subseteq \vdash_{\mathbf{AL_2}}$. □

As a result, the ALs discussed in Sections 5-7 are incomparable, i.e. an AL belonging to one of these three types cannot in general be stronger or weaker than an AL belonging to another of the three types.

8.2 *Prima facie* obligations revisited

Explosion principles To apply the criteria from Section 4.2 to the logics from Section 3.1, we need some more preparation. We take it that the premises of the explosion principles, resp. arguments under consideration are all concerned with *prima facie* obligations, whereas their conclusion concerns actual obligations. Under this translation, $\mathbf{SDL_p^r}$ and $\mathbf{SDL_p^m}$ invalidate the analogues of (DEX) and (DEX-$\mathsf{O}\bot$):

$$\mathsf{O^p}A \wedge \mathsf{O^p}\neg A \vdash \mathsf{O}B \tag{74}$$

$$\mathsf{O^p}(A \wedge \neg A) \vdash \mathsf{O}B \tag{75}$$

The other explosion principles cannot as easily be translated to these systems, because in Section 3.1 we did not define a corresponding *prima facie* permission operator for the logics $\mathbf{SDL_p^x}$.

Suppose that we add a second dummy operator $\mathsf{P^p}$ to the language of $\mathbf{SDL_p}$. For the adaptive extension of the resulting logic, we re-define the set of abnormalities Ω_p by including both formulas of the form $\mathsf{O^p}A \wedge \neg\mathsf{O}A$ *and* formulas of the form $\mathsf{P^p}A \wedge \neg\mathsf{P}A$. In the resulting logic, the following analogues of the explosion principles (DEX-$\mathsf{P}\bot$) and (DEX-$\mathsf{OP}\neg$) are invalid:

$$\mathsf{P^p}(A \wedge \neg A) \vdash \mathsf{O}B \tag{76}$$

$$\mathsf{O^p}A \wedge \mathsf{P^p}\neg A \vdash B \tag{77}$$

$$\mathsf{O^p}A \wedge \neg\mathsf{P^p}A \vdash B \tag{78}$$

Note that conflicts of the form $\mathsf{O^p}A \wedge \mathsf{P^p}\neg A$ give rise to disjunctions of abnormalities in this logic:

$$\mathsf{O^p}A \wedge \mathsf{P^p}\neg A \vdash (\mathsf{O^p}A \wedge \neg\mathsf{O}A) \vee (\mathsf{P^p}\neg A \wedge \neg\mathsf{P}\neg A) \tag{79}$$

In case there is a conflict between a *prima facie* obligation and a *prima facie* permission, the adaptive logic will not prioritize one over the other. This is in line with [?], where it is argued that permission should not take priority over obligations or conversely. Should one nevertheless want a logic that does treat one type of conflict as “worse” than the other, then one can turn to the format of lexicographic ALs as sketched in Section 3.4.

Benchmark examples First, in both $\mathbf{SDL^r_p}$ and $\mathbf{SDL^m_p}$, the Smith and Jones arguments are $\mathbf{SDL^x_p}$-valid, while Roberts and Thomas are not.

$$\mathsf{O^p}(f \vee s), \mathsf{O^p}\neg f \vdash_{\mathbf{SDL^x_p}} \mathsf{O}s \tag{Smith}$$

$$\mathsf{O^p}(j \wedge s) \vdash_{\mathbf{SDL^x_p}} \mathsf{O}j \tag{Jones}$$

$$\mathsf{O^p}(t \wedge r), \mathsf{O^p}(\neg t \wedge v) \nvdash_{\mathbf{SDL^x_p}} \mathsf{O}r \wedge \mathsf{O}v \tag{Roberts 1}$$

$$\mathsf{O^p}(t \wedge r), \mathsf{O^p}(\neg t \wedge v) \nvdash_{\mathbf{SDL^x_p}} \mathsf{O}(r \wedge v) \tag{Roberts 2}$$

$$\mathsf{O^p}(t \wedge (f \vee s)), \mathsf{O^p}(\neg t \wedge \neg f) \nvdash_{\mathbf{SDL^x_p}} \mathsf{O}s \tag{Thomas}$$

In order to infer the conclusions of the Roberts and Thomas arguments, we would need to detach the obligations $\mathsf{O}(t \wedge r)$ and $\mathsf{O}(t \wedge (f \vee s))$ respectively. But we cannot do that in view of the following minimal Dab-consequences of the respective premise sets:

$$(\mathsf{O^p}(t \wedge r) \wedge \neg\mathsf{O}(t \wedge r)) \vee (\mathsf{O^p}(\neg t \wedge v) \wedge \neg\mathsf{O}(\neg t \wedge v)) \tag{80}$$

$$(\mathsf{O^p}(t \wedge (f \vee s)) \wedge \neg\mathsf{O}(t \wedge (f \vee s))) \vee (\mathsf{O^p}(\neg t \wedge \neg f) \wedge \neg\mathsf{O}(\neg t \wedge \neg f)) \tag{81}$$

One way of accounting for the Roberts and Thomas arguments is to strengthen $\mathbf{SDL^x_p}$ by closing the operator $\mathsf{O^p}$ under a number of further rules. For instance, we could add a principle permitting the inference from $\mathsf{O^p}(A \wedge B)$ to $\mathsf{O^p}A$, such as (Inh). That would enable us to infer $\mathsf{O^p}r$ given $\mathsf{O^p}(t \wedge r)$, and $\mathsf{O}r$ given $\mathsf{O^p}r$ (on the condition $\mathsf{O^p}r \wedge \neg\mathsf{O}r$). Clearly, however, not anything goes when closing $\mathsf{O^p}$ under additional rules. For one thing, we do not want to end up with full **SDL** or even **K** for *prima facie* obligations, as this would completely annihilate our initial objective. But also if we characterize $\mathsf{O^p}$ in terms of weaker logics like the ones presented in Sections 5-7, we should be careful. After all, the richer one’s lower limit logic, the more likely one is to end up with flip-flop problems that will require further tinkering with the set of abnormalities, much as we had to do in previous sections.

For the Natascha argument, one can translate the impossibility of $s \wedge m$ using the operator O for actual obligations. The underlying idea is that constraints concerning what is practically (im)possible only have a bearing on actual obligations, not on the *prima facie* obligations. This can again be done in two different ways, giving rise to two different premise sets. For both, the validity of the argument will depend on the adaptive strategy:

$$\mathsf{O}^{\mathsf{p}}s, \mathsf{O}^{\mathsf{p}}m, \mathsf{O}^{\mathsf{p}}(s \supset t), \mathsf{O}^{\mathsf{p}}(m \supset t), \neg\mathsf{O}(s \wedge m) \nvdash_{\mathbf{SDL}^{\mathbf{r}}_{\mathbf{p}}} \mathsf{O}t \tag{Natascha 1}$$

$$\mathsf{O}^{\mathsf{p}}s, \mathsf{O}^{\mathsf{p}}m, \mathsf{O}^{\mathsf{p}}(s \supset t), \mathsf{O}^{\mathsf{p}}(m \supset t), \neg\mathsf{O}(s \wedge m) \vdash_{\mathbf{SDL}^{\mathbf{m}}_{\mathbf{p}}} \mathsf{O}t \tag{Natascha 1}$$

$$\mathsf{O}^{\mathsf{p}}s, \mathsf{O}^{\mathsf{p}}m, \mathsf{O}^{\mathsf{p}}(s \supset t), \mathsf{O}^{\mathsf{p}}(m \supset t), \mathsf{O}\neg(s \wedge m) \nvdash_{\mathbf{SDL}^{\mathbf{r}}_{\mathbf{p}}} \mathsf{O}t \tag{Natascha 2}$$

$$\mathsf{O}^{\mathsf{p}}s, \mathsf{O}^{\mathsf{p}}m, \mathsf{O}^{\mathsf{p}}(s \supset t), \mathsf{O}^{\mathsf{p}}(m \supset t), \mathsf{O}\neg(s \wedge m) \vdash_{\mathbf{SDL}^{\mathbf{m}}_{\mathbf{p}}} \mathsf{O}t \tag{Natascha 2}$$

In Sections 4–8 we defined and discussed a large variety of conflict-tolerant deontic logics that can be developed within the AL framework. More variation is possible, as there are other ways still to define conflict-tolerant deontic logics – by moving to a hyperintensional framework, for instance – and strengthen them adaptively. Moreover, existing systems can be altered by making them more expressive, e.g. by considering the interplay between deontic modalities and alethic, doxastic, or epistemic modalities. All this goes to show that adaptive logics provide a versatile and modular framework for conflict-tolerant normative reasoning, and that their applications to this problem are far from exhausted.

9 Conditional obligations and adaptive detachment

SDL is inadequate not just for accommodating normative conflicts in deontic logic, but also for representing deontic conditionals, as we will explain below.[59] Within the vast literature on such conditionals, one can distinguish three general approaches. The first is to represent them by means of a *dyadic* obligation operator $\mathsf{O}(\cdot \mid \cdot)$, and to read a formula $\mathsf{O}(B \mid A)$ as 'If A, then B is obligatory'. A second approach is to treat the problems surrounding deontic conditionals as symptomatic of the bigger challenge of how to formalize conditional statements in general. The third approach is more abstract: it treats deontic conditionals as pairs connecting a given "input" with an "output", and defines specific proof theories and an operational semantics (based on the principle of detachment and **CL**) for such connections.

We will discuss these three different approaches in Sections 9.1-9.3 respectively, showing how the framework of ALs can be useful in each of them. Our discussion will be mainly tentative; we provide pointers to more technical results and fully worked-out proposals in the literature at the end of each subsection.

[59] We will only sketch the latter inadequacy here. It is discussed at length in Section 8.5. and in the Appendix of [48, Chapter 1]. For other overviews of this problem, see for instance [3; 28].

9.1 Adaptive dyadic deontic logics

Helping one's neighbours Let us illustrate the distinctive problems surrounding deontic conditionals by means of a so-called Chisholm scenario – after [30]. This scenario can be represented as follows in the dyadic setting:

(i) It is obligatory that Jones goes to the aid of his neighbours ($\mathsf{O}g$).

(ii) It is obligatory that if Jones goes to the aid of his neighbours, then he tells them he is coming ($\mathsf{O}(t \mid g)$).

(iii) If Jones does not go to the aid of his neighbours, then he ought not to tell them he is coming ($\mathsf{O}(\neg t \mid \neg g)$).

(iv) Jones does not go to the aid of his neighbours ($\neg g$).

Recall now the principles of *factual detachment* (FD) and *deontic detachment* (DD) from Section 1:

$$A, \mathsf{O}(B \mid A) \vdash \mathsf{O}B \qquad \text{(FD)}$$

$$\mathsf{O}A, \mathsf{O}(B \mid A) \vdash \mathsf{O}B \qquad \text{(DD)}$$

Given premises (iii) and (iv), we can use (FD) to infer an obligation $\mathsf{O}\neg t$ for Jones not to tell his neighbours he is coming. However, given premises (i) and (ii), we can also use (DD) to infer an obligation $\mathsf{O}t$ for Jones to tell his neighbours he is coming.

But now we face a dilemma. Jones cannot both tell and not tell his neighbours he is coming. So, each of (DD) and (FD) has some intuitive appeal, but together they lead to a deontic conflict, and hence explosion if the logic of O is **SDL**. This is the dilemma of deontic and factual detachment, also known in the literature as "the dilemma of detachment and commitment" [3; 106]. In fact, one should rather speak here of a *trilemma*, since one may deny that **SDL** is an appropriate logic for obligations, and insist that both (FD) and (DD) should be unconditionally valid. This means one needs a conflict-tolerant deontic logic for O, much as those discussed in preceding sections. Here, we will first focus on the other two horns of the trilemma and exclude conflicts at the level of O.

Since each of (DD) and (FD) seems reasonable in isolation, Hilpinen and McNamara argue that we cannot just pick one of them at the expense of the other, and that we need to move to a more nuanced position beyond this choice [48, p. 119]. One solution is to make the detachment – via (DD) or (FD) – of unconditional obligations subject to further conditions, such as joint consistency. The AL framework allows us to make this idea exact, and to study its pros and cons.

A simple solution Let $\mathbf{SDL_d}$ be the logic obtained by replacing the unary *prima facie* operator $\mathsf{O}^{\mathsf{p}}(\cdot)$ of $\mathbf{SDL_p}$ with the conditional operator $\mathsf{O}(\cdot \mid \cdot)$. As we did with the O^{p}-operator of $\mathbf{SDL_p}$, we treat the new conditional operator like a dummy operator in $\mathbf{SDL_d}$.

Some authors treat unconditional obligations $\mathsf{O}A$ on the same foot as conditional obligations of the type $\mathsf{O}(A \mid \top)$. Note that in $\mathbf{SDL_d}$ these are not equivalent. For instance, the conjunction $\mathsf{O}(A \mid \top) \wedge \mathsf{O}(\neg A \mid \top)$ is $\mathbf{SDL_d}$-consistent, while the conjunction $\mathsf{O}A \wedge \mathsf{O}\neg A$ is not. In line with the interpretation in Section 3, $\mathsf{O}(A \mid \top)$ expresses something like "A is an unconditional prima facie obligation", whereas the intended reading of $\mathsf{O}A$ is that "A is an actual obligation".

In order to detach unconditional obligations from conditional obligations, we strengthen $\mathbf{SDL_d}$ adaptively to the logics $\mathbf{SDL_d^x}$, which are defined by the triple $\langle \mathbf{SDL_d}, \Omega_d, \mathsf{x}\rangle$, with $\mathsf{x} \in \{\mathsf{r}, \mathsf{m}\}$ and $\Omega_d = \Omega_{fd} \cup \Omega_{dd}$:

$$\Omega_{fd} = \{\mathsf{O}(B \mid A) \wedge A \wedge \neg\mathsf{O}B \mid A, B \in \mathcal{W}\}$$
$$\Omega_{dd} = \{\mathsf{O}(B \mid A) \wedge \mathsf{O}A \wedge \neg\mathsf{O}B \mid A, B \in \mathcal{W}\}$$

In view of the $\mathbf{SDL_d}$-valid inferences (82) and (83), the adaptive logics $\mathbf{SDL_d^x}$ allow for the conditional application of (FD) and (DD):

$$A, \mathsf{O}(B \mid A) \vdash \mathsf{O}B \vee (\mathsf{O}(B \mid A) \wedge A \wedge \neg\mathsf{O}B) \tag{82}$$
$$\mathsf{O}A, \mathsf{O}(B \mid A) \vdash \mathsf{O}B \vee (\mathsf{O}(B \mid A) \wedge \mathsf{O}A \wedge \neg\mathsf{O}B) \tag{83}$$

We illustrate the resulting logic by applying it to the Chisholm scenario in (i)-(iv):

1	$\mathsf{O}g$	Prem	$\emptyset$
2	$\mathsf{O}(t \mid g)$	Prem	$\emptyset$
3	$\mathsf{O}(\neg t \mid \neg g)$	Prem	$\emptyset$
4	$\neg g$	Prem	$\emptyset$
5	$\mathsf{O}t$	1,2; RC	$\{\mathsf{O}(t \mid g) \wedge \mathsf{O}g \wedge \neg\mathsf{O}t\}\checkmark^7$
6	$\mathsf{O}\neg t$	3,4; RC	$\{\mathsf{O}(\neg t \mid \neg g) \wedge \neg g \wedge \neg\mathsf{O}\neg t\}\checkmark^7$
7	$(\mathsf{O}(t \mid g) \wedge \mathsf{O}g \wedge \neg\mathsf{O}t)\vee$ $(\mathsf{O}(\neg t \mid \neg g) \wedge \neg g \wedge \neg\mathsf{O}\neg t)$	1-4; RU	$\emptyset$

Lines 4 and 5 remain marked in any extension of this proof, so that neither $\mathsf{O}t$ nor $\mathsf{O}\neg t$ is an $\mathbf{SDL_d^x}$-consequence of the premises at lines 1-4. Thus, in cases of conflict, the applications of (FD) and (DD) that lead to the conflict are rejected.

Some have taken a bolder stance here by arguing that when factual and deontic detachment lead to a conflict, (FD) overrules (DD) or vice versa. We will not go into this discussion here – see [48, p. 112-124] for an overview of the various positions.

However, let us briefly indicate how this idea of overruling can be modeled with the AL framework.

Recall the lexicographic ALs that were introduced in Section 3.4. Consider the lexicographic ALs defined in terms of the lower limit logic $\mathbf{SDL_d}$ and the sequence $\langle\Omega_{fd}, \Omega_{dd}\rangle$. The idea is that we treat abnormalities with respect to factual detachment as "worst", and hence give priority to (FD) over (DD). For instance, in the Chisholm case, the abnormality $\mathsf{O}(\neg t \mid \neg g) \wedge \neg g \wedge \neg\mathsf{O}\neg t$ will be avoided, and hence the abnormality $\mathsf{O}(t \mid g) \wedge \mathsf{O}g \wedge \neg\mathsf{O}t$ will be assumed to hold. Thus, in such logics, one can conclude that Jones ought not to tell his neighbours he is coming. Other applications of (DD) that do not result in conflicting obligations will remain valid in such logics. Finally, if two different applications of (FD) conflict, they will both be blocked in the adaptive logics.

A (prioritized) combination of various sorts of adaptive reasoning may also be useful for those who insist on the intuitiveness of (FD) and (DD), and use these to cast doubt on the validity of full **SDL** for O (cf. our discussion of the trilemma of detachment and commitment, supra). Here, one may combine insights and techniques from Sections 5–7 with those from the present section, treating each of (FD), (DD), and (some or all) rules and axioms of **SDL** as defeasible. This way one cannot only accommodate deontic conflicts that arise from an applications of either (FD) or (DD) or both – by invalidating those applications – but also conflicting obligations that happen to be simply there, "unconditionally". In such a setting, one may e.g. prioritize the standard behavior of O over the applicability of (FD) and (DD), thus capturing the intuition that even if they are sometimes to be accepted, deontic conflicts should be avoided whenever possible.

Open problems and further reading The first monotonic dyadic deontic logics were introduced in Bengt Hansson's seminal paper [46; 75]. Hansson-style dyadic deontic logics typically invalidate (FD), while some of them validate (DD).

More recently, van Benthem, Grossi and Liu have investigated the relation between modal logics of preferences, priority structures, and dyadic deontic logic more generally [97]. In this account, the factual information in the antecedent of (FD) is formalized as a dynamic epistemic event, rather than as a "mere" factual (propositional) statement. This way, the non-monotonicity of reasoning with dyadic obligations is formalized at the object-level, rather than as a property of the consequence relation.

Our focus in this section was on the defeasible application of the detachment principles (FD) and (DD), in a language with both a dyadic operator $\mathsf{O}(\cdot \mid \cdot)$ for conditional obligations and an independent, monadic operator O that satisfies full **SDL**. We did not discuss other logical properties of $\mathsf{O}(\cdot \mid \cdot)$, and instead treated it as a dummy operator much like we treated the O^{p}-operator from Section 3. But we may of course wonder whether there are no logical properties which the dyadic operator ought to satisfy unrestrictedly. Possible candidates include, for instance, the dyadic versions of the aggregation and inheritance principles:

$$(\mathsf{O}(B \mid A) \wedge \mathsf{O}(C \mid A)) \supset (\mathsf{O}(B \wedge C \mid A)) \tag{DAgg}$$

$$\text{From } \mathsf{O}(B \mid A) \text{ and } \vdash B \supset C, \text{ to infer } \mathsf{O}(C \mid A) \tag{DInh}$$

However, one has to be careful again, since enriching one's lower limit logic may easily give rise to flip-flop-problems, analogous to the monadic deontic logics presented in previous sections. The solutions that were discussed in those sections may in turn be transferred to the dyadic setting.

Different preferences regarding the characterization of $\mathsf{O}(\cdot \mid \cdot)$ have given rise to a wide variety of dyadic systems, including a range of conflict-tolerant dyadic systems which could in turn be extended adaptively so as to gain further inferential power. For instance, in [90] and [87, Ch. 11], Christian Straßer studied conditional versions of some of the **LUM**-systems from Section 5, and presented a number of adaptive extensions of these logics. In [91] and [87, Chapters 11–12], Straßer presents a general method for turning dyadic deontic logics into ALs which allow for the conditional application of (FD), paying special attention to Chisholm-scenarios.

Finally, it should also be noted that, even if we leave (FD) and (DD) aside, all the observations and techniques from Sections 5–7 could be applied just as well to the case of dyadic deontic logics as developed, building on Goble's work in [39; 40]. Here again, we may use adaptive logics to steer a middle course between all-too-weak conflict tolerant dyadic systems and deontic explosion.

9.2 Adaptive reasoning with conditionals

Adaptive detachment, generalized Instead of using a binary operator for conditional obligation, one may also introduce a new conditional $\Rightarrow$, so that the logic of deontic conditionals derives from the logic for this new conditional and the logic for the monadic operator O of one's choice. In this section we focus on this second approach.

Suppose we formalize "If A, then B is obligatory" as $A \Rightarrow \mathsf{O}B$.[60] Then at the very least we want to be able to factually detach $\mathsf{O}B$ given A and $A \Rightarrow \mathsf{O}B$, *absent further information.*[61] But we may not want unrestricted detachment (or full modus

[60] One may also represent the conditional obligation "If A, then it is obligatory that B" by $\mathsf{O}(A \Rightarrow B)$ or $\mathsf{O}A \Rightarrow \mathsf{O}B$. We will have little to say about the first of these two alternatives; we briefly return to the second at the end of this section.

[61] We consider deontic detachment at the end of this section.

ponens) for the conditional $\Rightarrow$. For instance, given the premises $p, q, p \Rightarrow \mathsf{O}r$, and $q \Rightarrow \mathsf{O}\neg r$, we may not want to be able to detach both $\mathsf{O}r$ and $\mathsf{O}\neg r$, unless perhaps we move to a non-standard characterization of O. So if we stick to a standard characterization of O as an **SDL**-operator, we will want to allow for some, but not all instances of modus ponens for $\Rightarrow$.

In other words, we only want to apply detachment in a defeasible way. This can be done as follows in terms of ALs. We first enrich the language of **SDL** with a default conditional, where nested occurrences of $\Rightarrow$ are disallowed:

$$\begin{aligned} \mathcal{W}^{\Rightarrow} := \quad & \mathcal{W}^{d} \mid \langle \mathcal{W}^{d} \rangle \Rightarrow \langle \mathcal{W}^{d} \rangle \mid \neg\langle \mathcal{W}^{\Rightarrow} \rangle \mid \langle \mathcal{W}^{\Rightarrow} \rangle \vee \langle \mathcal{W}^{\Rightarrow} \rangle \mid \langle \mathcal{W}^{\Rightarrow} \rangle \wedge \\ & \langle \mathcal{W}^{\Rightarrow} \rangle \mid \\ & \langle \mathcal{W}^{\Rightarrow} \rangle \supset \langle \mathcal{W}^{\Rightarrow} \rangle \mid \langle \mathcal{W}^{\Rightarrow} \rangle \equiv \langle \mathcal{W}^{\Rightarrow} \rangle \end{aligned}$$

Next, let $\mathbf{SDL}_{\Rightarrow}$ be just **SDL**, but defined over this richer language. Hence, $\Rightarrow$ has no properties in $\mathbf{SDL}_{\Rightarrow}$. We then define our ALs on the basis of $\mathbf{SDL}_{\Rightarrow}$, by the set of abnormalities

$$\Omega_{\Rightarrow} =_{\mathsf{df}} \{(A \Rightarrow B) \wedge A \wedge \neg B \mid A, B \in \mathcal{W}^{d}\}$$

So whenever the conditional $A \Rightarrow B$ is true and A is true, then we assume that also B is true. Note that A and B can be arbitrary members of $\mathcal{W}^{d}$, hence also A can be a deontic statement such as $\mathsf{O}p$ – we return to this point below.

Let us call the resulting adaptive logics $\mathbf{SDL}^{\mathbf{x}}_{\Rightarrow}$. As the following proof illustrates, conditional obligations are detachable in $\mathbf{SDL}^{\mathbf{x}}_{\Rightarrow}$ as long as no conflicts are generated. (For the sake of readability, we abbreviate $(A \Rightarrow B) \wedge A \wedge \neg B$ as $A \not\Rightarrow B$).)

1	$p \wedge q$	Prem	$\emptyset$
2	$p \Rightarrow \mathsf{O}r$	Prem	$\emptyset$
3	$q \Rightarrow \mathsf{O}\neg r$	Prem	$\emptyset$
4	$(p \wedge q) \Rightarrow \mathsf{O}s$	Prem	$\emptyset$
5	$\mathsf{O}r$	1,2;RC	$\{p \not\Rightarrow \mathsf{O}r\}\checkmark^{8}$
6	$\mathsf{O}\neg r$	1,3;RC	$\{q \not\Rightarrow \mathsf{O}\neg r\}\checkmark^{8}$
7	$\mathsf{O}s$	1,4;RC	$\{(p \wedge q) \not\Rightarrow \mathsf{O}s\}$
8	$(p \not\Rightarrow \mathsf{O}r) \vee (q \not\Rightarrow \mathsf{O}\neg r)$	1-3;RU	$\emptyset$

The conditional $\Rightarrow$ of $\mathbf{SDL}_{\Rightarrow}$ is of course very weak – we can only make use of it by going adaptive. We can however strengthen the lower limit logic by adding further rules. Here are some candidates:

$$\text{If } A \Rightarrow C \text{ and } B \Rightarrow C\text{, then } (A \vee B) \Rightarrow C \tag{Or}$$

$$\text{If } A \Rightarrow B \text{ and } B \Rightarrow C\text{, then } A \Rightarrow C \tag{Tra}$$

$$\text{If } A \Rightarrow B \text{ and } (A \wedge B) \Rightarrow C\text{, then } A \Rightarrow C \tag{CTra}$$

$$\text{If } A \vdash B \text{ and } B \Rightarrow C\text{, then } A \Rightarrow C \tag{SA}$$

Each of these rules can be added to our logic if desired. However, one should be careful here, as adding more properties to one's lower limit logic often generates flip-flop problems, as explained in the previous sections of this paper.

Unlike the dyadic deontic operator of $\mathbf{SDL_d}$ from Section 9.1, the conditional $\Rightarrow$ of $\mathbf{SDL}^{\mathbf{x}}_{\Rightarrow}$ is completely independent of the way we formalize obligations. We can read a statement $A \Rightarrow B$ as 'If A, then normally B' as we would do for defeasible conditionals in general. In $\mathbf{SDL}^{\mathbf{x}}_{\Rightarrow}$ we detach obligations via defeasible modus ponens, just like we defeasibly detach conclusions in default logic or in your preferred calculus of non-monotonic logic. So this approach is very unifying, treating deontic reasoning as just one specific type of defeasible reasoning in general.

However, the approach has the disadvantage that it cannot as easily accommodate deontic detachment (DD) (cf. Section 9.1). Consider the following three inferences:

$$p, p \Rightarrow \mathsf{O}q \vdash \mathsf{O}q \tag{84}$$

$$\mathsf{O}p, \mathsf{O}p \Rightarrow \mathsf{O}q \vdash \mathsf{O}q \tag{85}$$

$$\mathsf{O}p, p \Rightarrow \mathsf{O}q \vdash \mathsf{O}q \tag{86}$$

(84) and (85) are derivable $\mathbf{SDL}^{\mathbf{x}}_{\Rightarrow}$-rules: we can apply these rules conditionally in $\mathbf{SDL}^{\mathbf{x}}_{\Rightarrow}$. However, (86) is not a derivable rule in $\mathbf{SDL}^{\mathbf{x}}_{\Rightarrow}$. Some have argued that this is how it should be (see e.g. the discussion and references in [25]). Still, (86) has some intuitive force.

One way to defend $\mathbf{SDL}^{\mathbf{x}}_{\Rightarrow}$ is by arguing that, whenever we think deontic detachment should be allowed, the appropriate translation of the conditional is as in (85). More generally, such conditionals are of the form: if A is obligatory, then also B is obligatory ($\mathsf{O}A \Rightarrow \mathsf{O}B$). However, that would mean that in many cases we need a kind of "double translation" of deontic conditionals – as $(A \Rightarrow \mathsf{O}B) \wedge (\mathsf{O}A \Rightarrow \mathsf{O}B)$ – which seems highly artificial. Moreover, it would go against the spirit of the adaptive logic approach, where the idea is that the logic should determine which applications of deontic detachment are rational. So altogether, it seems that the second approach is less suited to accommodate (DD).

Further reading The literature on the formalization of defeasible conditionals is vast. For some good entry points, see e.g. [55; 61]. In this section we only presented a basic mechanism for the defeasible detachment of obligations via a new conditional. For more information on the types of rules that can be studied via this mechanism, we refer to [87, Chapter 6].

9.3 Adaptive Characterizations of input/output logic

Input/output logic The third approach to deontic conditionals that we will discuss here goes under the name input/output logic (henceforth I/O logic). Technically speaking, I/O logics (without constraints, cf. infra) are operations that map every pair $\langle \mathcal{A}, \mathcal{G} \rangle$ to an "output" $\mathcal{O} \subseteq \mathcal{W}$, where (i) $\mathcal{G} \subseteq \mathcal{W} \times \mathcal{W}$ is a set of "input/output pairs" (A, B); (ii) $\mathcal{A} \subseteq \mathcal{W}$ is the "input". For instance, given the input $\mathcal{A} = \{p, q\}$ and the set of conditionals $\mathcal{G} = \{(p, r), (q, s)\}$, the output $\mathcal{O}$ will consist of r, s, and everything that follows from their conjunction.

In a deontic setting, $\mathcal{A}$ usually represents factual information, $\mathcal{G}$ is a set of conditional obligations, and the output consists of what is obligatory, given the facts at hand and given the conditional obligations that make up our normative system. The idea of factual detachment thus lies at the very core of I/O-logics.

Different I/O-logics are obtained by varying on the rules under which $\mathcal{G}$ is closed, before one applies factual detachment. These rules are themselves highly similar to the ones used to characterize default conditionals (cf. Section 9.2). For example, by assuming that $\mathcal{G}$ is closed under the rule (OR)

$$\text{If } (A, C) \text{ and } (B, C), \text{ then } (A \vee B, C) \tag{OR}$$

we can obtain r in the output of $\mathcal{A} = \{p \vee q\}$ and $\mathcal{G} = \{(p, r), (q, r)\}$. Similarly, if $\mathcal{G}$ is closed under the rule (Tra), one can validate deontic detachment (DD):

$$\text{If } (A, B) \text{ and } (B, C), \text{ then } (A, C) \tag{Tra}$$

So for instance, given closure under (Tra), we can obtain q in the output of $\mathcal{A} = \emptyset$ and $\mathcal{G} = \{(\top, p), (p, q)\}$.

Both (FD) and (DD) are accommodated within the I/O-systems presented [58]. However, this framework cannot handle conflicts that arise from the application of (FD) or (DD) or both: e.g. $\mathcal{A} = \{p, q\}$ and $\mathcal{G} = \{(p, r), (q, \neg r)\}$ will generate a trivial output.

To deal with such cases, Makinson and van der Torre introduced a set $\mathcal{C}$ of "constraints" in their [59]. Depending on the application context $\mathcal{C}$ may represent physical constraints, human rights, practical considerations, etc. $\mathcal{C}$ can restrict the output in two ways, each corresponding to a different style of reasoning. We can require consistency of $\mathcal{O} \cup \mathcal{C}$, or we can impose the weaker requirement that for each $A \in \mathcal{O}$, $\{A\} \cup \mathcal{C}$ is consistent. In the border case where $\mathcal{C} = \emptyset$, this simply means that we require the $\mathcal{O}$ to be consistent, or that each $A \in \mathcal{O}$ is consistent. The first approach is called *meet* constrained output; the second is the *join* constrained output.

The adaptive characterization In [88], I/O-logics are characterized in terms of deductive systems within a rich modal language. We explain how this works for constrained I/O-logics (the case for unconstrained I/O-logics is simpler). The language uses unary modal operators $\mathsf{in}, \mathsf{out}, \mathsf{con}$ to represent input, output, and constraints respectively. Input/output pairs (A, B) are represented by means of $\mathsf{in}, \mathsf{out}$ and a conditional $\rightarrow$, as follows:

$$\mathsf{in}A \rightarrow \mathsf{out}B$$

The principle of detachment and the rules for input/output-pairs are then translated into the object level. This gives us rules and axioms such as the following:

$$\text{If } \mathsf{in}A \text{ and } \mathsf{in}A \rightarrow \mathsf{out}B\text{, then } \mathsf{out}B \qquad (\text{DET}')$$

$$((\mathsf{in}A \rightarrow \mathsf{out}C) \wedge (\mathsf{in}B \rightarrow \mathsf{out}C)) \supset (\mathsf{in}(A \vee B) \rightarrow \mathsf{out}C) \qquad (\text{OR}')$$

$$((\mathsf{in}A \rightarrow \mathsf{out}B) \wedge (\mathsf{in}B \rightarrow \mathsf{out}C)) \supset (\mathsf{in}A \rightarrow \mathsf{out}C) \qquad (\text{Tra}')$$

The fact that the output should be consistent with the set of constraints is captured by

$$\mathsf{con}A \supset \neg\mathsf{out}\neg A \qquad (\text{ROC})$$

Finally, to mimic the selection of maximal consistent sets of conditionals, a dummy operator $\bullet$ is introduced and used in much the same way as we did in Section 3. That is, conditionals $(A, B) \in \mathcal{G}$ are translated into formulas of the form $\bullet(\mathsf{in}A \rightarrow \mathsf{out}B)$. The adaptive logics then allow one to "activate" such conditionals by removing the dummy, whence one can apply rules like (DET′), (OR′), or (Tra′) to them.

Suppose, for instance, that we are given the following set of inputs, I/O-pairs, and constraints: $\mathcal{A} = \{p, q\}, \mathcal{G} = \{(p, r), (q, s), (p, t)\}, \mathcal{C} = \{\neg r \vee \neg s\}$. In the language from [88], this gives us the following premise set:

$$\Gamma = \{\mathsf{in}p, \mathsf{in}q, \bullet(\mathsf{in}p \rightarrow \mathsf{out}r), \bullet(\mathsf{in}q \rightarrow \mathsf{out}s), \bullet(\mathsf{in}p \rightarrow \mathsf{out}t), \mathsf{con}(\neg r \vee \neg s)\}$$

In an adaptive proof from Γ, we can finally derive $\mathsf{out}t$. Depending on the strategy, we can also finally derive $\mathsf{out}(r \vee s)$ or even $\mathsf{out}r$ and $\mathsf{out}s$.

Let us illustrate this with an object-level proof. To enhance readibility, we use $\star(A, B)$ to abbreviate $\bullet(\mathsf{in}A \to \mathsf{out}B) \wedge \neg(\mathsf{in}A \to \mathsf{out}B)$. Moreover, we use superscripts r, m to indicate the strategy under which certain lines are (not) marked:[62]

1	$\mathsf{in}p$	Prem	$\emptyset$
2	$\mathsf{in}q$	Prem	$\emptyset$
3	$\bullet(\mathsf{in}p \to \mathsf{out}r)$	Prem	$\emptyset$
4	$\bullet(\mathsf{in}q \to \mathsf{out}s)$	Prem	$\emptyset$
5	$\bullet(\mathsf{in}p \to \mathsf{out}t)$	Prem	$\emptyset$
6	$\mathsf{con}(\neg r \vee \neg s)$	Prem	$\emptyset$
7	$\mathsf{in}p \to \mathsf{out}r$	3; RC	$\{\star(p, r)\}\checkmark^{\mathsf{r,m}}$
8	$\mathsf{in}q \to \mathsf{out}s$	4; RC	$\{\star(q, s)\}\checkmark^{\mathsf{r,m}}$
9	$\mathsf{in}p \to \mathsf{out}t$	5; RC	$\{\star(p, t)\}$
10	$\mathsf{out}r$	1,7; RU	$\{\star(p, r)\}\checkmark^{\mathsf{r,m}}$
11	$\mathsf{out}s$	2,8; RU)	$\{\star(q, s)\}\checkmark^{\mathsf{r,m}}$
12	$\mathsf{out}t$	1,9; RU	$\{\star(p, t)\}$
13	$\mathsf{out}r \vee \mathsf{out}s$	10; RU	$\{\star(p, r)\}\checkmark^{\mathsf{r}}$
14	$\mathsf{out}r \vee \mathsf{out}s$	11; RU	$\{\star(q, s)\}\checkmark^{\mathsf{r}}$
15	$\star(p, r) \vee \star(q, s)$	1-4,6; RU	$\emptyset$

Under the modal translation, the minimal abnormality strategy corresponds to the operation of meet constrained output; normal selections (cf. Section 3.4 corresponds to the join constrained output. The reliability strategy has no counterpart in the original framework of [59]; however, as shown in [88], one can also define a procedural semantics for the corresponding operation, much in the spirit of Makinson and van der Torre's original setting.

Further reading I/O-logic was introduced by Makinson and van der Torre [58; 59] as a formal tool for modeling non-monotonic reasoning with conditionals. We refer to [76] for an introduction to this approach and its applications to deontic reasoning.

The framework presented here is not only sufficient to characterize many well-known I/O logics, but it allows one to go beyond the expressive means of I/O logics so as to express useful notions in deontic logic such as violations and sanctions. We refer to [88] for the many details, and for an elaborate presentation and discussion of these advantages.

[62] The formulas at lines 10-12 are derivable in view of (DET′). The formula at line 15 is derivable in view of (DET′), modal properties of the **KD**-operator out, and the axiom schema (ROC).

10 Deontic compatibility

10.1 Adaptive logics for deontic compatibility

We saw how ALs are useful for reasoning in the presence of normative conflicts, and for detaching conditional obligations. A different context of application for ALs that was mentioned in Section 1 concerns the implementation of the *nullum crimen sine lege* principle (henceforth NCSL). This principle expresses that no crimes occur where there is no law: that which is not forbidden, is permitted. Typically, NCSL is understood as a rule of closure permitting all the actions not prohibited by penal law [1, pp. 142–143]. It is a fundamental principle of law, the roots of which go back at least as far as the French Revolution. In the twentieth century it was incorporated in various human rights instruments as a non-derogable right [70].

Logicians and computer scientists are very familiar with the concept of "negation by default", according to which a piece of information represented by some variable is taken to be absent unless and until we include it in our database. For instance, where a variable x abbreviates that there is a train leaving for Ghent at 14:14, we may conclude that $\neg x$ unless x is mentioned on the timetable at the train station. Similarly, we can think of NCSL as "permission by default". Formally, this can be expressed as follows, where we take our premise set Γ to represent a given normative system or law, and where $\vdash$ is an ordinary (Tarskian) deontic logic:

$$\Gamma \vdash \mathsf{P}A \text{ iff } \Gamma \nvdash \neg\mathsf{P}A$$

Assume that we want to implement this equivalence against the background of full **SDL**. Then, on pain of inconsistency, the equivalence can at best hold *defeasibly.* Suppose, for instance, that we are given a premise set Γ such that $\Gamma \vdash \neg\mathsf{P}p \vee \neg\mathsf{P}q$, while $\Gamma \nvdash \neg\mathsf{P}p$ and $\Gamma \nvdash \neg\mathsf{P}q$. Then we cannot preserve consistency *and* apply NCSL to derive $\mathsf{P}p$ as well as $\mathsf{P}q$. What we want, then, is a logic that preserves consistency and applies NCSL *as much as possible.*

This motivates an adaptive logic of deontic compatibility which implements NCSL by taking **SDL** as its lower limit logic, and Ω_{P} as its set of abnormalities:

$$\Omega_{\mathsf{P}} = \{\neg\mathsf{P}A \mid A \in \mathcal{W}\}$$

We call the resulting logic $\mathbf{SDL}^{\mathsf{x}}_{\mathbf{nc}}$ with nc for *nullum crimen* and $\mathsf{x} \in \{\mathsf{r},\mathsf{m}\}$. In view of the **SDL**-validity of $\mathsf{P}A \vee \neg\mathsf{P}A$, $\mathbf{SDL}^{\mathsf{x}}_{\mathbf{nc}}$ allows for the inference of jointly compatible permissions relative to a given premise set. The following object level proof further illustrates the ways this logic works.

1	$\mathsf{O}(\neg p \vee \neg q)$	Prem	$\emptyset$
2	$\mathsf{O}(\neg s \wedge t)$	Prem	$\emptyset$
3	$\mathsf{P}t \supset (\mathsf{P}u \supset \mathsf{O}\neg v)$	Prem	$\emptyset$
4	$\mathsf{P}p$	RC	$\{\neg\mathsf{P}p\}$
5	$\mathsf{P}\neg p$	RC	$\{\neg\mathsf{P}\neg p\}$
6	$\mathsf{P}q$	RC	$\{\neg\mathsf{P}q\}$
7	$\mathsf{P}\neg q$	RC	$\{\neg\mathsf{P}\neg q\}$
8	$\mathsf{P}r$	RC	$\{\neg\mathsf{P}r\}$
9	$\mathsf{P}\neg r$	RC	$\{\neg\mathsf{P}\neg r\}$
10	$\mathsf{P}s$	RC	$\{\neg\mathsf{P}s\}\checkmark^{18}$
11	$\mathsf{P}\neg s$	2; RU	$\emptyset$
12	$\mathsf{P}t$	2; RU	$\emptyset$
13	$\mathsf{P}\neg t$	RC	$\{\neg\mathsf{P}\neg t\}\checkmark^{19}$
14	$\mathsf{P}u$	RC	$\{\neg\mathsf{P}u\}\checkmark^{20}$
15	$\mathsf{P}\neg u$	RC	$\{\neg\mathsf{P}\neg u\}$
16	$\mathsf{P}v$	RC	$\{\neg\mathsf{P}v\}\checkmark^{20}$
17	$\mathsf{P}\neg v$	RC	$\{\neg\mathsf{P}\neg v\}$
18	$\neg\mathsf{P}s$	2; RU	$\emptyset$
19	$\neg\mathsf{P}\neg t$	2;RU	$\emptyset$
20	$\neg\mathsf{P}u \vee \neg\mathsf{P}v$	2;3;RU	$\emptyset$

One nice feature of this logic is its simplicity, when restricted to premise sets of the form $\{\mathsf{O}A \mid A \in \Delta\}$ for $\Delta \subseteq \mathcal{W}$. Indeed, for such cases, the strategies *reliability* and *minimal abnormality* will coincide, since every minimal Dab-consequence of such premise sets contains only one disjunct $A \in \Omega_{\mathsf{P}}$. This is itself an immediate corollary of the following:

Proposition 10.1. *If* $\Gamma = \{\mathsf{O}A \mid A \in \Delta\}$ *for* $\Delta \subseteq \mathcal{W}$*, then* $\Gamma \vdash_{\mathbf{SDL}} (\neg\mathsf{P}A_1 \vee \ldots \vee \neg\mathsf{P}A_n)$ *iff there is an* $i \in \{1, \ldots, n\}$ *such that* $\Gamma \vdash_{\mathbf{SDL}} \neg\mathsf{P}A_i$*.*

In more complex cases such as our example proof above, the two strategies may well differ. In either case, the resulting consequence set will be closed under **SDL** and consistent.

One may wonder whether the idea of deontic compatibility should necessarily be phrased in terms of the underlying logic **SDL** – after all, legal conflicts are a fact of life, and as soon as such conflicts are modeled in **SDL**, everything becomes obligatory and permissible. This motivates a logic that defeasibly applies NCSL *and* that accommodates conflicts much as the logics presented in Sections 5-7.

Let us illustrate this by means of the paraconsistent deontic logics from Section 7. One option is to just take a monotonic paraconsistent deontic logic – say **DCLuN**, to keep things relatively simple – and to use as a set of abnormalities

$$\Omega = \{\mathsf{O}A \mid A \in \mathcal{W}^{\sim}\}$$

However, the resulting logic will be too strong, in the sense that it will allow one to derive permissions that should intuitively not be derivable, even if we take NCSL seriously. With such a logic, one can e.g. derive $\mathsf{P}^{\urcorner}_{\neg}{\sim}p$ from $\Gamma = \{\mathsf{O}p\}$. The underlying reason is that in these logics, $\mathsf{O}p$ does not entail $\mathsf{O}\neg{\sim}p$ (just like the truth of p does not entail the falsehood of ${\sim}p$ in their paraconsistent propositional base), and hence one can consistently assume that $\mathsf{O}\neg{\sim}p$ is false even when $\mathsf{O}p$ is true. But the mere fact that we want to allow for the logical possibility of conflicts, should not entail that everything is permissible.

A more plausible combination of conflict-tolerance and *nullum crimen* can be obtained if we combine the *adaptive* logics $\mathbf{DCLuN}^{\mathsf{x}}$ from Section 7 with NCSL, using the format of lexicographic ALs that was introduced in Section 3.4. This means that the logic first minimizes inconsistencies (which implies i.a. that we derive further obligations), and only after that do we maximize permissions. In this way we can e.g. explain why in view of $\Gamma' = \{\mathsf{O}p, \mathsf{O}({\sim}p \vee q), \mathsf{O}r, \mathsf{O}{\sim}r\}$ we can derive $\mathsf{O}q$, $\mathsf{O}p$ and $\neg\mathsf{P}^{\urcorner}_{\neg}{\sim}p$, $\neg\mathsf{P}^{\urcorner}_{\neg}{\sim}q$, but also $\mathsf{P}^{\urcorner}_{\neg}s, \mathsf{P}^{\urcorner}_{\neg}{\sim}s$, and $\mathsf{P}^{\urcorner}_{\neg}r, \mathsf{P}^{\urcorner}_{\neg}{\sim}r$.

Analogously, one may enrich the logics from Sections 5 and 6 with a default version of NCSL. For similar reasons as in the paraconsistent case, it seems best to first apply the adaptive mechanisms from those sections, and only after that to apply NCSL. For instance, in the case of non-aggregative deontic logics, we would not want to infer $\mathsf{P}\neg(p \wedge q)$ from $\Gamma = \{\mathsf{O}p, \mathsf{O}q\}$. Likewise, in the context of the **LUM**-logics, we would not want to infer $\mathsf{P}\neg p$ from $\Gamma' = \{\mathsf{O}(p \wedge q)\}$. The full development of such rich ALs for deontic compatibility is still very much open; it should by now be clear that a broad range of options are to be considered, and that the devil may well be in the many details.

10.2 Further reading

Adaptive logics for classical compatibility were among the first ampliative adaptive logics to be published – see [13]. Although these logics were not formulated in the standard format, one can do this by means of the triple

$$\langle \mathbf{S5}, \{\neg\Diamond A \mid A \text{ is a non-modal formula }\}, \mathsf{x} \in \{\mathsf{r}, \mathsf{m}\}\rangle$$

The relation between classical compatibility and the logics in question is then expressed in terms of a modal translation: A is compatible with Γ iff $\{\Box A \mid A \in \Gamma\} \vdash_{\mathbf{AL}} \Diamond A$.

In [65], the basic idea behind these logics is used in order to develop a formal account of paraconsistent compatibility, i.e., what it means that a given formula is compatible with a certain (possibly inconsistent) scientific theory. As Meheus argues there, one also first needs to minimize inconsistencies before checking compatibility with the resulting maximally consistent interpretation of the theory.

11 Summary and outlook

This paper started with two simple adaptive logics that can handle deontic conflicts. We then discussed in some detail more sophisticated conflict-tolerant ALs, as well as ALs for reasoning with conditional obligations and the problems of detachment that are associated with these. Finally, we broadened the picture by presenting ALs for the inherently defeasible *nullum crimen sine lege* principle. This should convince the reader of the generality and the flexibility of the adaptive logic framework.

It is important to realize, however, that this does not exhaust the possibilities of adaptive logics for the domain of normative reasoning. This requires more explanation.

All logics presented in this paper share important constraints. One of them is that we only considered the two main deontic modalities, “it is obligatory that” and “it is permitted that”, and we moreover restricted our formal languages to non-nested occurrences of those modalities. Another one is that we took it for granted that we can start from premise sets that merely consist of very specific and very concrete normative statements, like “Nathan ought to take Lisa to that particular movie on Saturday afternoon”.

Because of these constraints, the logics allow us to explicate only a very small part of the normative reasoning one finds in actual cases. Already the everyday examples from Nathan’s life (that are recognizable to many of us) suffice to illustrate this. In Nathan’s first predicament (the preludium), his normative reasoning does not start from the statements that he ought to take Lisa to the movie in the afternoon, that he ought to look after Ben in the afternoon and that he ought to take Lisa for a veggie burger in the evening. These statements are themselves *derived* from other statements, in this case concrete promises by Nathan and the general rule “One ought to keep one’s promises”. Also in Nathan’s second predicament (Section 3.1), the specific normative statements are not given at the outset, but are the result of reasoning. In this case, not only general rules play a role (like “One ought to return favors”), but also commands uttered by an authority (i.c. Nathan’s father). None of the logics presented here allows us to explicate the reasoning from general rules to their instances or from commands (uttered by one person) to obligations (for another person) – to mention only two possible origins of specific normative statements.

There is more. Some readers may have noticed that, while presenting our conflict-tolerant logics, we used the term "*prima facie* obligations", but never used the term "all-things-considered obligations" which is, at least since Ross' [83], associated with it. Instead we consistently used the term "actual obligations". The reason is that none of our logics enables us to explicate the reasoning from *prima facie* obligations to all-things-considered obligations, where the latter is taken to mean something like "obligations that are, after careful deliberation, considered to be binding". Our logics only give us those binding obligations for which relatively little deliberation is needed. For instance, "if a *prima facie* obligation is unconflicted, it should be binding" or "if two *prima facie* obligations are unconflicted, also their conjunction should be binding", etc.

In order to explicate the reasoning that goes on in resolving a predicament and finding out what one's all-things-considered obligations are (or should be), we need much more than just deontic operators. For instance, whatever Nathan's solution for his first predicament may be, it will involve certain beliefs (for instance, what Nathan believes will happen if he does not keep the promise he made to his mother). None of our logics can handle interactions between deontic modalities on the one hand, and doxastic or epistemic modalities on the other.[63]

Does this mean we have gone all this way for nothing? Certainly not. We are convinced that the logics presented here are good candidates to explicate *part of* the reasoning that goes on in specific deontic contexts. They moreover provide a first stepping stone to more complex, richer accounts of deontic reasoning. So there is still hope for Nathan, or at least for us to fully understand how he should reason.

References

[1] Carlos E. Alchourrón and Eugenio Bulygin. *Normative Systems.* Springer-Verlag, Wien/New York, 1971.

[2] Patrick Allo. Adaptive logic as a modal logic. *Studia Logica*, 101(5):933–958, 2013.

[3] Lennart Åqvist. Deontic logic. In Dov Gabbay and Franz Guenthner, editors, *Handbook of Philosophical Logic (2nd edition)*, volume 8, pages 147–264. Kluwer Academic Publishers, 2002.

[4] D. Batens. Dialectical dynamics within formal logics. *Logique et Analyse*, 114:161–173, 1986.

[63] See e.g. [73] for a study of the interaction between epistemic and deontic modalities.

[5] D. Batens. Dynamic dialectical logics. In G. Priest, R. Routley, and J. Norman, editors, *Paraconsistent Logic. Essays on the Inconsistent*, pages 187–217. Philosophia Verlag, München, 1989.

[6] Diderik Batens. Inconsistencies and beyond. A logical-philosophical discussion. *Revue Internationale de Philosophie*, 200:259–273, 1997.

[7] Diderik Batens. Inconsistency-adaptive logics. In Ewa Orłowska, editor, *Logic at Work. Essays dedicated to the memory of Helena Rasiowa*, pages 445–472. Physica Verlag (Springer), Heidelberg, New York, 1999.

[8] Diderik Batens. Zero logic adding up to classical logic. *Logical Studies*, 2:15, 1999.

[9] Diderik Batens. A general characterization of adaptive logics. *Logique et Analyse*, 173–175:45–68, 2001. Appeared 2003.

[10] Diderik Batens. A universal logic approach to adaptive logics. *Logica Universalis*, 1:221–242, 2007.

[11] Diderik Batens. Logics for qualitative inductive generalization. *Studia Logica*, 97:61–80, 2011.

[12] Diderik Batens. Tutorial on inconsistency-adaptive logics. In Jean-Yves Béziau, Mihir Chakraborty, and Soma Dutta, editors, *Springer Proceedings in Mathematics & Statistics*, volume 152, pages 3–38. Springer, 2015.

[13] Diderik Batens and Joke Meheus. The adaptive logic of compatibility. *Studia Logica*, 66:327–348, 2000.

[14] Diderik Batens, Christian Straßer, and Peter Verdée. On the transparency of defeasible logics: Equivalent premise sets, equivalence of their extensions, and maximality of the lower limit. *Logique et Analyse*, 207:281–304, 2009.

[15] M. Beirlaen. *Tolerating Normative Conflicts in Deontic Logic.* Dissertation, Ghent University, 2012. Available online at `http://www.clps.ugent.be/research/doctoral-dissertations`.

[16] M. Beirlaen and C. Straßer. Two adaptive logics of norm-propositions. *Journal of Applied Logic*, 11(2):147–168, 2013.

[17] M. Beirlaen and C. Straßer. Nonmonotonic reasoning with normative conflicts in multi-agent deontic logic. *Journal of Logic and Computation*, 24:1179–1207, 2014.

[18] M. Beirlaen and C. Straßer. A structured argumentation framework for detaching conditional obligations. In O. Roy, A. Tamminga, and M. Willer, editors, *Proceedings of the 13th International Conference on Deontic Logic and Normative Systems (ΔEON 2016, Bayreuth, Germany)*, pages 32–48. College Publications, 2016.

[19] Mathieu Beirlaen and Atocha Aliseda. A conditional logic for abduction. *Synthese*, 191(15):3733–3758, 2014.

[20] Mathieu Beirlaen, Bert Leuridan, and Frederik Van De Putte. A logic for the discovery of deterministic causal regularities. *Synthese*, 195:367–399, 2018.

[21] Mathieu Beirlaen and Christian Straßer. A paraconsistent multi-agent framework for dealing with normative conflicts. In Joao Leite, Paolo Torroni, Thomas Agotnes, Guido Boella, and Leon van der Torre, editors, *Computational Logic in Multi-Agent*

Systems, volume 6814 of *Lecture Notes in Computer Science*, pages 312–329. Springer, Berlin/Heidelberg, 2011.

[22] Mathieu Beirlaen, Christian Straßer, and Joke Meheus. An inconsistency-adaptive deontic logic for normative conflicts. *Journal of Philosophical Logic*, 42(2):285–315, 2013.

[23] N. Belnap and M. Perloff. In the realm of agents. *Annals of Mathematics and Artificial Intelligence*, 9:25–48, 1993.

[24] Patrick Blackburn, Maarten De Rijke, and Yde Venema. *Modal Logic.* Cambridge Tracts in Theoretical Computer Science, 2001.

[25] Daniel Bonevac. Defaulting on reasons. *Noûs*, 2016.

[26] D. Brink. Moral conflict and its structure. *The Philosophical Review*, 103:215–247, 1994.

[27] Fabrizio Cariani. “Ought” and resolution semantics. *Noûs*, 47(3):534–558, 2013.

[28] J. Carmo and A. Jones. Deontic logic and contrary-to-duties. In Dov Gabbay and Franz Guenthner, editors, *Handbook of Philosophical Logic (2nd edition)*, volume 8, pages 265–343. Kluwer Academic Publishers, 2002.

[29] Brian Chellas. *Modal Logic: an Introduction.* Cambridge: Cambridge university press, 1980.

[30] R. Chisholm. Contrary-to-duty imperatives and deontic logic. *Analysis*, 27:33–36, 1963.

[31] Dag Elgesem. The modal logic of agency. *Nordic Journal of Philosophical Logic*, 2(2):1–46, 1997.

[32] Kit Fine. Angellic content. *Journal of Philosophical Logic*, 45(2):199–226, April 2016.

[33] Dov M. Gabbay. Bipolar argumentation frames and contrary to duty obligations, preliminary report. In M. Fisher, L. van der Torre, M. Dastani, and G. Governatori, editors, *Computational Logic in Multi-Agent Systems*, pages 1–24. Springer, 2012.

[34] L. Goble. A logic of “good”, “should”, and “would”: Part I. *Journal of Philosophical Logic*, 19:169–199, 1990.

[35] L. Goble. A proposal for dealing with deontic dilemmas. In A. Lomuscio and D. Nute, editors, *7th International Workshop on Deontic Logic in Computer Science*, volume 3065 of *Lecture Notes in Computer Science*, pages 74–113. Springer, 2004.

[36] L. Goble. Normative conflicts and the logic of ought. *Noûs*, 43:450–489, 2009.

[37] Lou Goble. A logic of “good”, “should”, and “would”: Part II. *Journal of Philosophical Logic*, 19:253–76, 1990.

[38] Lou Goble. Multiplex semantics for deontic logic. *Nordic Journal of Philosophical Logic*, 5:113–134, 2000.

[39] Lou Goble. Preference semantics for deontic logic. Part I: Simple models. *Logique et Analyse*, 183–184:383–418, 2003.

[40] Lou Goble. Preference semantics for deontic logic. Part II: Multiplex models. *Logique et Analyse*, 185–188:335–363, 2004.

[41] Lou Goble. A logic for deontic dilemmas. *Journal of Applied Logic*, 3:461–483, 2005.

[42] Lou Goble. Prima facie norms, normative conflicts, and dilemmas. In Dov Gabbay, Leon van der Torre, John Horty, and Xavier Parent, editors, *Handbook of Deontic Logic and Normative Systems*, volume 1, chapter 4, pages 241–351. College Publications, 2013.

[43] Lou Goble. Deontic logic (adapted) for normative conflicts. *Logic Journal of the IGPL*, 22(2):206–235, 2014.

[44] C.W. Gowans, editor. *Moral Dilemmas.* Oxford University Press, 1987.

[45] J. Hansen. Imperative logic and its problems. In Dov Gabbay, Leon van der Torre, John Horty, and Xavier Parent, editors, *Handbook of Deontic Logic and Normative Systems*, volume 1, chapter 2, pages 137–192. College Publications, 2013.

[46] Bengt Hansson. An analysis of some deontic logics. *Nous*, 3:373–398, 1969.

[47] Jesse Heyninck and Christian Straßer. Relations between assumption-based approaches in nonmonotonic logic and formal argumentation. In Gabriele Kern-Isberner and Renata Wassermann, editors, *16th International Workshop on Non-Monotonic Reasoning, Cape Town, South Africa*, pages 65–76, 2016.

[48] Risto Hilpinen and Paul McNamara. Deontic logic: a historical survey and introduction. In Dov Gabbay, Leon van der Torre, John Horty, and Xavier Parent, editors, *Handbook of Deontic Logic and Normative Systems*, volume 1, chapter 1, pages 3–136. College Publications, 2013.

[49] J. Horty. Moral dilemmas and nonmonotonic logic. *Journal of Philosophical Logic*, 23(1):35–66, 1994.

[50] J. Horty. Nonmonotonic foundations for deontic logic. In Donald Nute, editor, *Defeasible Deontic Logic: Essays in Nonmonotonic Normative Reasoning*, pages 17–44. Kluwer Academic Publishers, 1997.

[51] J. Horty. Skepticism and floating conclusions. *Artificial Intelligence*, 135:55–72, 2002.

[52] J. Horty. Reasoning with moral conflicts. *Noûs*, 37:557–605, 2003.

[53] J. Horty. *Reasons as Defaults.* Oxford University Press, 2012.

[54] Frank Jackson. On the semantics and logic of obligation. *Mind*, 94:177–195, 1985.

[55] S. Kraus, D. Lehmann, and M. Magidor. Nonmonotonic reasoning, preferential models and cumulative logics. *Artificial Intelligence*, 44:167–207, 1990.

[56] Bert Leuridan. Causal discovery and the problem of ignorance. An adaptive logic approach. *Journal of Applied Logic*, 7(2):188–205, 2009.

[57] D. Makinson and K. Schlechta. Floating conclusions and zombie paths: two deep difficulties in the “directly skeptical” approach to defeasible inheritance nets. *Artificial Intelligence*, 48:199–209, 1991.

[58] D. Makinson and L. van der Torre. Input/output logics. *Journal of Philosophical Logic*, 29:383–408, 2000.

[59] D. Makinson and L. van der Torre. Constraints for input/output logics. *Journal of Philosophical Logic*, 30:155–185, 2001.

[60] David Makinson. General patterns in nonmonotonic reasoning. In *Handbook of Logic in Artificial Intelligence and Logic Programming, vol. III*. Clarendon Press, 1994.

[61] David Makinson. *Bridges from Classical to Nonmonotonic Logic*, volume 5 of *Texts in Computing*. King's College Publications, London, 2005.

[62] Ruth Barcan Marcus. Moral dilemmas and consistency. *Journal of Philosophy*, 77:121–136, 1980. Reprinted in [44].

[63] C. McGinnis. *Paraconsistency and Deontic Logic: Formal Systems for Reasoning with Normative Conflicts*. Dissertation, University of Minnesota, 2007.

[64] C. McGinnis. Semi-paraconsistent deontic logic. In Jean-Yves Béziau, Walter Carnielli, and Dov Gabbay, editors, *Handbook of Paraconsistency*, pages 81–99. College Publications, London, 2007.

[65] Joke Meheus. Paraconsistent compatibility. *Logique et Analyse*, 183–184:251–287, 2003.

[66] Joke Meheus, Mathieu Beirlaen, and Frederik Van De Putte. Avoiding deontic explosion by contextually restricting aggregation. In Guido Governatori and Giovanni Sartor, editors, *Deontic Logic in Computer Science*, volume 6181 of *Lecture Notes in Computer Science*, pages 148–165. Springer Berlin Heidelberg, 2010.

[67] Joke Meheus, Mathieu Beirlaen, Frederik Van De Putte, and Christian Straßer. Non-adjunctive deontic logics that validate aggregation as much as possible. Unpublished manuscript, 2012. Preprint available at `http://www.clps.ugent.be/research/publications`.

[68] Joke Meheus, Christian Straßer, and Peter Verdée. Which style of reasoning to choose in the face of conflicting information? *Journal of Logic and Computation*, 26(1):361–380, 2016.

[69] Joke Meheus, Liza Verhoeven, Maarten Van Dyck, and Dagmar Provijn. Ampliative adaptive logics and the foundation of logic-based approaches to abduction. In L. Magnani, N.J. Nersessian, and Claudio Pizzi, editors, *Logical and Computational Aspects of Model-Based Reasoning*, pages 39–71. Kluwer Academic, Dordrecht, 2002.

[70] Ali Mokhtar. Nullum crimen, nulla poena sine lege: Aspects and prospects. *Statute Law Review*, 26(1):41–55, 2005.

[71] D. Nute. Norms, priorities, and defeasibility. In Paul McNamara and Henri Prakken, editors, *Norms, Logics and Information Systems. New Studies on Deontic Logic and Computer Science.*, pages 201–218. IOS Press, 1999.

[72] Sergei Odintsov and Stanislav Speranski. Computability issues for adaptive logics in expanded standard format. *Studia Logica*, 101(6):1237–1262, 2013.

[73] Eric Pacuit, Rohit Parikh, and Eva Cogan. The logic of knowledge based obligation. *Synthese*, 149(2):311–341, 2006.

[74] X. Parent and L. van der Torre. "Sing and dance!" Input/output logics without weakening. In F. Cariani, D. Grossi, J. Meheus, and X. Parent, editors, *DEON (12th International Conference on Deontic Logic in Computer Science)*, volume 8554 of *Lecture Notes in Artificial Intelligence*, pages 149–165. Springer, 2014.

[75] Xavier Parent. A complete axiom set for Hansson's deontic logic DSDL2. *Logic Journal of the IGPL*, 18(3):422–429, 2010.

[76] Xavier Parent and Leendert van der Torre. Input/output logic. In Dov Gabbay, Jeff Horty, Xavier Parent, Ron van der Meyden, and Leendert van der Torre, editors, *Handbook of Deontic Logic and Normative Systems*, volume 1, chapter 8, pages 499–544. College Publications, 2013.

[77] H. Prakken and G. Sartor. Law and logic: A review from an argumentation perspective. *Artificial Intelligence*, 227:214–245, 2015.

[78] Henri Prakken. Intuitions and the modelling of defeasible reasoning: some case studies. In *Proceedings of the Ninth International Workshop on Nonmonotonic Reasoning*, pages 91–99, Toulouse, 2002.

[79] G. Priest. Paraconsistent logic. In D. Gabbay and F. Guenthner, editors, *Handbook of Philosophical Logic (2nd edition)*, volume 8, pages 287–393. Kluwer Academic Publishers, 2002.

[80] G. Priest, K. Tanaka, and Z. Weber. Paraconsistent logic. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Spring 2015 edition, 2015. http://plato.stanford.edu/archives/spr2015/entries/logic-paraconsistent/.

[81] Graham Priest. *In Contradiction. A Study of the Transconsistent.* Nijhoff, Dordrecht, 1987.

[82] N. Rescher and R. Manor. On inferences from inconsistent premises. *Theory and Decision*, 1:179–217, 1970.

[83] W. David Ross. *The Right and the Good.* Clarendon Press, 1930.

[84] Peter K. Schotch and Raymond E. Jennings. Non-kripkean deontic logic. In Risto Hilpinen, editor, *New Studies in Deontic Logic*, pages 149–162. Reidel, Dordrecht, 1981.

[85] K. Segerberg. An essay in classical modal logic, 1971.

[86] K. Segerberg. Getting started: beginnings in the logic of action. *Studia Logica*, 51:347–378, 1992.

[87] C. Straßer. *Adaptive Logic and Defeasible Reasoning. Applications in Argumentation, Normative Reasoning and Default Reasoning.* Springer, 2014.

[88] C. Straßer, M. Beirlaen, and F. Van De Putte. Dynamic proof theories for input/output logic. *Studia Logica*, 104:869–916, 2016.

[89] C. Straßer, A. Knoks, and Joke Meheus. Deontic reasoning on the basis of consistency considerations. Under review, 2017.

[90] Christian Straßer. An adaptive logic framework for conditional obligations and deontic dilemmas. *Logic and Logical Philosophy*, 19(1-2):95–128, 2010.

[91] Christian Straßer. A deontic logic framework allowing for factual detachment. *Journal of Applied Logic*, 9:61–80, 2011.

[92] Christian Straßer and Ofer Arieli. Normative reasoning by sequent-based argumentation. *Journal of Logic and Computation*, 2015 (online first).

[93] Christian Straßer and Mathieu Beirlaen. Towards more conflict-tolerant deontic

logics by relaxing the interdefinability between obligations and permissions. Unpublished manuscript. Preprint available at `http://www.clps.ugent.be/research/publications`.

[94] Christian Straßer, Joke Meheus, and Mathieu Beirlaen. Tolerating deontic conflicts by adaptively restricting inheritance. *Logique et Analyse*, 219:477–506, 2012.

[95] Christian Straßer and Dunja Šešelja. Towards the Proof-theoretic Unification of Dung's Argumentation Framework: an Adaptive Logic Approach. *Journal of Logic and Computation*, 21:133–156, 2010.

[96] Johan van Benthem. What one may come to know. *Analysis*, 64(282):95–105, 2004.

[97] Johan van Benthem, Davide Grossi, and Fenrong Liu. Priority structures in deontic logic. *Theoria*, 80(2):116–152, 2014.

[98] Frederik Van De Putte. *Generic Formats for Prioritized Adaptive Logics. With Applications in Deontic Logic, Abduction and Belief Revision.* Dissertation, Ghent University, 2012. Available at `http://www.clps.ugent.be/research/doctoral-dissertations`.

[99] Frederik Van De Putte. Default assumptions and selection functions: A generic framework for non-monotonic logics. In Felix Castro, Alexander Gelbukh, and Miguel Gonzalez, editors, *Advances in Artificial Intelligence and Its Applications*, volume 8265 of *Lecture Notes in Computer Science*, pages 54–67. Springer, 2013.

[100] Frederik Van De Putte. Coarse Deontic Logic. In Allard Tamminga and Malte Willer, editors, *Deontic Logic and Normative Systems: 13th International Conference, DEON 2016, Bayreuth, Germany*, pages 256–271. College Publications, July 2016.

[101] Frederik Van De Putte. Coarse Deontic Logic (Extended Version). *Journal of Logic and Computation*, 29(2):285–317, 2019.

[102] Frederik Van De Putte and Christian Straßer. Extending the standard format of adaptive logics to the prioritized case. *Logique et Analyse*, 220:601–641, 2012.

[103] Frederik Van De Putte and Christian Straßer. A logic for prioritized normative reasoning. *Journal of Logic and Computation*, 23(3):563–583, 2013.

[104] Frederik Van De Putte and Christian Straßer. Adaptive logics: a parametric approach. *Logic Journal of IGPL*, 22(6):905–932, 2014.

[105] L. van der Torre and S. Villata. An ASPIC-based legal argumentation framework for deontic reasoning. In *Computational Models of Argument (Proceedings of COMMA 14)*, pages 421–432. IOS Press, 2014.

[106] J.A. van Eck. A system of temporally relative modal and deontic predicate logic and its philosophical applications. *Logique et Analyse*, 99:249–290, 1982.

[107] Bas C. van Fraassen. Values and the heart's command. *Journal of Philosophy*, 70(1):5–19, 1973.

[108] Peter Verdée. Non-monotonic set theory as a pragmatic foundation of mathematics. *Foundations of Science*, 18(4):655–680, Nov 2013.

[109] Georg Henrik von Wright. Deontic logic. *Mind*, 60:1–15, 1951.

[110] P. Vranas. I ought, therefore I can. *Philosophical Studies*, 136:167–216, 2007.

[111] Bernard Williams and W.F̃. Atkinson. Symposium: Ethical consistency. *Proceedings of the Aristotelian Society, Supplementary Volumes*, 39:103–138, 1965.

Received 13 June 2018