If, Then, Ought: From Logic to Computation

Xavier Parent

Mémoire présenté en vue de l'obtention de
l'Habilitation à diriger des recherches
Spécialité informatique
Université de Toulouse, 118 route de Narbonne, 31062 Toulouse cedex 9

A. Herzig	Research Director, CNRS-IRIT	Coordinateur
P. Cabalar	Professor, Corunna University	Rapporteur
S. Cranefield	Professor, Otago University	Examineur
H. van Ditmarsch	Research Director, CNRS-IRIT	Examineur
O. Gasquet	Professor, University of Toulouse	Examineur
C. Strasser	Professor, Bochum University	Rapporteur
I. Varzinczak	Professor, Sorbonne Paris North University	Rapporteur

Contents

Chapter 1

Introduction

1.1 Thesis summary

Norms pervade various aspects of life, and deontic logic delves into their logical analysis, revealing their contextual nature in the form of conditional statements.

In this habilitation thesis, I showcase my contributions to the field of deontic logic, spanning over the past 20 years, with a primary focus on the logic of conditional norms. My contributions aim to address two fundamental gaps in the area, and they are thus organized along two axes, a main one devoted to axiomatization (with two sub-axes), and a secondary one devoted to automation.

The first axe revolves around the elucidation of axiomatization problems, with a particular emphasis on two prevailing paradigms: the modal logic paradigm (sub-axe 1) and the norm-based (or rule-based) paradigm (sub-axe 2).[a] The norm-based paradigm treats the normative system as a first-class citizen, and originates in the work on conflict-tolerant deontic logics, see e.g. [114]. Deontic modalities are analyzed not in terms of possible worlds, but with respect to a set of explicitly defined norms−whether regulative, such as obligations and permissions, or constitutive, which are rules that define or create the very possibility of certain actions or institutions (e.g., "a valid contract requires mutual consent"). This approach moves away from a truth-functional semantics in favor of an operational one, where detachment (modus ponens) serves as the core mechanism. The central question becomes:

[a]The term "norm-based" was first introduced by Hansen [48] and has since entered common usage in the field. The term "rule-based" is also occasionally used.

given a certain input, which obligations can be detached? This avoids some of the contentious assumptions of truth-functional semantics, such as treating norms as bearing a truth-value or as being based on a maximization principle. Moreover, a norm-based semantics allows for a more precise control over detached obligations and better handles concepts that challenge traditional modal logic, such as explicit permission and conflicts between obligations.

A gap existed in the absence of a roadmap detailing the main systems within each of these paradigms. In order to bridge this gap, I have concentrated my efforts on two well-established representatives from each paradigm. Both of these representatives are inherently more complex than SDL [25], a modal logic of type KD, which is known to be unsuitable for normative reasoning, because of the deontic paradoxes (like Chisholm [26]'s paradox).

First, I have made contributions to the family of systems known as preference-based dyadic deontic logics, which have their origins in the works of Hansson, Lewis, and others. Makinson [65] characterizes them as a non-monotonic formalism ahead of its time ("avant la lettre"). The main building block is a conditional obligation operator $\bigcirc(B/A)$, read as "If A is the case, then B is obligatory". A betterness relation ranks the possible worlds in terms of comparative goodness. The evaluation rule is in terms of best antecedent-worlds. It puts $\bigcirc(B/A)$ as true, when all the best A-worlds are B-worlds.

I have contributed a systematic axiomatic study, the first of its kind, encompassing all the various systems that can be obtained, based on (i) the properties of the betterness relation in the models and (ii) the concept of "best" employed when determining the truth-value of a conditional obligation. The lack of an analog of the modal logic cube was a gap in the field. Early completeness results were for the case where the betterness relation comes with the full panoply of the standard properties, resulting in the collapse of the different notions of "best". My central contribution revolved around the creation of this roadmap, thereby providing a solution to problems initially posed by Åqvist in his 1987 book [2], requiring innovative methods. Noteworthy discoveries emerged during this journey, such as the idleness of totality and weakened forms of transitivity, which are subjects of discussion in economics. The field still has unresolved issues that need to be addressed. The extension to the first-order case, which is crucial for ethical reasoning, is one of them.

The second family of systems I have engaged with (in collaboration with L. van der Torre) is input/output (I/O) logic, initially developed by my co-author in collaboration with D. Makinson. I/O logic is a well-established framework for normative

reasoning, as demonstrated by its dedicated chapter in the first volume of the *Handbook of Deontic Logic and Normative Systems* [92], as well as its inclusion in the *Stanford Encyclopedia of Philosophy* entry on deontic logic. I/O logic has primarily been used in deontic logic, normative multi-agent systems, and AI & Law. Recently, it has begun to attract interest within the knowledge representation and reasoning (KRR) field, particularly under the influence of Bochman [18]'s important work on causal reasoning. Bochman explicitly acknowledges that his inference system for causal rules "originates in input/output logic[]" [18, p. 17]. Although the recent study by Ciabattoni and Rozplokhas [31] on I/O logic received the Best Paper Award at KR 2023—highlighting the framework's growing recognition beyond deontic logic—it initially struggled to gain acceptance. Much of its broader visibility was due to Bochman's foundational contributions, which helped extend its relevance outside the traditional domain of deontic logic. More recently, the work by Arieli et al. [4] has further supported this shift by situating I/O logic within the broader landscape of formalisms commonly used in KRR, particularly default logic.

I/O logic displays the key characteristics associated with the norm-based systems outlined above. The specific mechanisms through which it achieves this will be discussed in greater detail in Chapter 3. Given that this framework may be less familiar to some readers than the other two introduced here, I will first provide a brief overview of its key ideas. I will then highlight the feature that distinguishes it from other norm-based approaches, such as Horty's default deontic logic [53] and Hansen's imperativist logic [47].

In this framework, a conditional obligation is represented as a pair of Boolean formulas (a, x), where a is the input (antecedent condition) and x is the output (normative consequence). A normative system N is defined as a set of such pairs. The semantics of I/O logic is defined procedurally: outputs are generated from given inputs according to a set of rules. The core expression is:

$$x \in out(N, a)$$

which is interpreted as: x is an output of the normative system N given input a. The role of the semantics is to define the operation *out*. The proof theory of I/O logic is formulated in terms of inference rules that operate on pairs of formulas, rather than on individual formulas, emphasizing the transformation of input-output pairs rather than the preservation of truth.

Soundness and completeness theorems show the equivalence between the syntactical and semantical characterizations. Thus, when a derivation contains a node labeled with (a, x), under the hood the semantics tells us that $x \in out(N, a)$.

There is a notable analogy between I/O logic and preference-based dyadic deontic logic, which can be illustrated as follows:

Preference-based:	$\bigcirc(x/a)$ valid in model M
I/O-based:	$x \in out(N, a)$

The distinctive feature of input/output logic is its rejection of the law of identity, expressed as "if a then a". In the I/O terminology, the input need not be in the output, and so the rule ID must go:

$$\frac{}{(a, a)}\ (\text{ID})$$

Makinson was initially motivated by the desire to eliminate the deontic counterpart of identity–"a, then it ought to be the case that a"–a principle that, at the time, was widely regarded as a problematic feature of deontic logic. Although validated in preference-based systems, its acceptance was highly contested and became a focal point of debate. Notably, the principle remains valid in Horty's deontic default logic and Hansen's imperativist semantics.

Together with L. van der Torre, I conducted the first systematic axiomatic study of the various systems that arise from different definitions of the input/output operation in the semantics of I/O logic. This investigation extended beyond the four traditional I/O logics originally defined by Makinson, leading to the discovery of an entire spectrum of weaker systems. To address the well-known paradoxes of contrary-to-duty (CTD) reasoning, I/O logic introduces a second layer known as constrained I/O logic. In this framework, a set of constraints is applied to filter or restrict the output, thereby eliminating undesirable obligations from the output set. Other norm-based semantic frameworks have restricted themselves to reasoning about norm conflicts. They typically lack the ability to handle contrary-to-duty (CTD) reasoning, viz. reasoning about norm violation. To bridge this gap, I developed a prioritized version of I/O logic able to do both simultaneously. Notably, constrained I/O logic previously lacked a proof theory. I addressed this by formulating a "surrogate" proof theory, offering a structured way to reason within this framework. This extended system has been applied to a significant problem in AI ethics: evaluating whether moral particularism can support a bottom-up approach to acquiring normative knowledge, as advocated by some scholars.

This work on axiomatization provided a solid foundation for my subsequent research, which focused on the mechanization of normative reasoning. This work, conducted

in collaboration with C. Benzmüller (University of Bamberg, Germany), was motivated by my intention to engage more seriously with computer science than I had previously. Automated Theorem Proving (ATP) is a rapidly evolving field, yet it has not been applied to conditional normative reasoning thus far. While there are some ATP systems available for SDL, there have been none for more complex frameworks, like preference-based dyadic deontic logic and input/output logic. To address this gap, we have developed a library of normative reasoners, following the shallow semantical embedding approach developed by my co-author and his team. The basic idea consists in faithfully embedding the target logics into higher-order logic (HOL) and then use an off-the-shelf HOL prover for automation. This indirect method offers a high degree of flexibility and has been successfully applied to a number of deontic logics, including the preference-based dyadic deontic logics studied in axe 1. We consider two possible uses of the framework. The first one is as a tool for meta-reasoning about the considered logics. The second use is as a tool for assessing ethical arguments. As a case study, we provide a computer encoding of a well-known paradox in population ethics, Parfit's repugnant conclusion.

1.2 Context and research environnment

Most of my research is in the area known as deontic logic, for which a standard reference is the *Handbook of Deontic Logic and Normative Systems*, whose two volumes I have co-edited [39, 40].

I am at the cross-road between computer science and philosophy. I believe in their fruitful cross-fertilization. This explains a lot of my research. I take most of my inspiration from philosophy, my primary area of study. Computer scientists rarely engage with philosophical literature, and so miss philosophers' insights and some perspectives. For instance, decades of philosophical debate have developed nuanced answers to, e.g., the trolley problem discussed in AI ethics. AI's moral dilemmas are not entirely new, so solutions should draw on philosophical progress. The more faithful computer science is to philosophy, the more trustworthy and nuanced it will be. Conversely, philosophers rarely engage with the computer science literature, which is also unfortunate. For example, to determine whether AI systems truly "think" or "reason" requires looking at the latest progress in machine learning and reasoning systems.

The publications range from 2008 to 2024, covering the period from 5 years after the PhD defense to the present. They cover three research projects I had in parallel. My interest in preference-based dyadic logic goes back to my PhD, defended under

the supervision of P. Livet. If it was not for him I would not be here. My interest in input/output logic arose during the ten years spent in Luxembourg, while that in automated reasoning arose during a visit by Benzmüller in Luxembourg.

My work in deontic logic was largely shaped by the research environment I was part of, even though my initial interests were in other areas like multi-agent systems and the philosophy of language. While those other topics were appealing to me, they didn't materialise in my work as much as deontic logic did. Perhaps the opportunities, resources, or collaborative possibilities in my environment naturally steered me in that direction.

I am deeply grateful to A. Herzig for agreeing to be my mentor. I was already in his debt after my PhD, since Andreas acted as the external reviewer—clearly, he enjoys keeping me on my toes!

1.3 Method of presentation

This thesis is organised in three chapters. Chapter 2 deals with my contributions to the axiomization problem of preference-based dyadic deontic logics. This work has been conducted as a single author except for the on-going work on first-order logic, which has been done with D. Pichler, whose PhD I co-supervise. Chapter 3 outlines my contributions to norm-based dyadic deontic logics, more specifically input/output logic. This work was mainly done in collaboration with L. van der Torre, with additional contributions from M. Olszewski, whose master I co-supervised. Chapter 4 presents my contributions to automated reasoning, done in collaboration with C. Benzmüller, with additional input from L. van der Torre, A. Farjami (whose PhD I co-supervised), and P. Medert (whose master I co-supervised). Chapter 5 highlights the lines of research I will pursue in the future. I distinguish between my short-term and long-term plans.

I was lucky to have many other very fruitful collaborations, but I have chosen not to present them. For instance, I have chosen not to include my collaboration with A. Jones at the King's College London immediately after the PhD on the formalization of speech acts theory and on the theory of normative positions.

Although I have occasionally cited standard works and relevant recent contributions in the field, I have intentionally kept these references to a minimum. The standard references are available in the papers I am presenting. The limited number of citations should not be interpreted as an attempt to overstate my role in these research areas; that is certainly not my intention.

When referring to work I conducted alone, I use "I". For joint work, I use "we", which refers to "the authors, including myself, of the publication being discussed". The specific individuals included in "we" are listed in the references.

The described papers are referenced in the main text, but also after each relevant section under "Publications". The papers included in the Habilitation are listed in the next subsection. The presentation extends beyond these papers to provide a broader perspective on my research.

1.4 List of included papers

I have selected six papers that I consider most significant—two from each of the three main research axes I have had. Selecting these papers was a challenging task, as I view research as a cumulative process–where the most recent work often feels like "your best paper ever." Instead of focusing solely on my latest publications, I have chosen those that mark key milestones in my research journey.

1. "Completeness of Aqvist's systems E and F", *Review of Symbolic Logic*, 2015, 8 (1), pp. 164-177.
2. "On some weakened forms of transitivity in the logic of conditional obligation", *Journal of Philosophical Logic*, 2024, 53, pp. 721–760.
3. "Moral particularism in the light of deontic logic", *Artificial Intelligence and Law* (DEON 10 special issue), 2011, 19 (2-3), pp. 75-98.
4. "I/O Logics with a consistency check" (with L. van der Torre). In J. Broersen et al. (eds.), *Deontic Logic and Normative Systems - 14th International Conference, DEON 2018*, 2018, College Publications, London, pp. 285-299.
5. "Designing normative theories of ethical reasoning" (with C. Benzmüller and L. van der Torre), *Artificial Intelligence*, 2020, 287, 103348.
6. "Conditional normative reasoning as a fragment of HOL" (with C. Benzmüller), *Journal of Applied Non-Classical Logic*, 2024, pp. 34(4), pp. 561–592.

Items 1 and 2 deal with preference-based dyadic deontic logic. A bridge is made between (deontic) logic and economics (rational choice theory). The bridge was suggested to me by Rott [102], who focused on belief change theory. Item 1 presents some completeness results for classes of preference models without relying on the limit assumption. It also covers the limiting case where the betterness relation in the models does not have any specific properties. Among the various results I obtained, this one was the least obvious, making the discovery of its proof especially rewarding. The paper appeared in 2015, and definitively confirmed one of the key

hypotheses that triggered my work. I first made this hypothesis in 2008, noting it in a footnote of my first paper on the subject [76]. I also discussed it with Åqvist during a coffee break at DEON 2008, though I lacked a formal proof at that time. The hypothesis posits that the distinction between different notions of "best" is largely inconsequential for the logic. Item 2 builds on the work reported in item 1, and pushes the boundaries further. It sheds fresh light on the later framework, bringing together various weakened forms of transitivity discussed in economics and exploring their role in deontic logic. The findings reported there were quite unexpected. This work made possible the application made in item 6 to population ethics−the first of its kind. This application is described in Section 4.4.

I chose item 3 because it likely best illustrates the relevance of my research to computer science for a broader audience. In this paper (as mentioned in Section 1) I define a prioritized variant of I/O logic. Normative reasoning–reasoning about obligations, permissions, and prohibitions—is fundamentally nonmonotonic. In a preference-based modal logic setting, the consequence relation is monotonic: once a conclusion is derived from a set of premises, adding more premises cannot invalidate that conclusion. However, it has long ago been recognized that normative reasoning often involves exceptions, conflicts, and context-dependent rules, where new information can overturn previous conclusions. I have found it a lot more easier to model nonmonotonic reasoning in a norm-based setting like I/O logic, especially when conflicts are resolved using a priority relation on obligations. In Section 1, I mentioned that I applied the framework to a key problem in AI ethics: assessing whether moral particularism can justify a bottom-up approach to acquiring normative knowledge, as proposed by some scholars. This is a retrospective interpretation of the paper: when the work was originally written, AI ethics had not yet gained the attention it has today, and I must say it is only six years later that I saw the connection. Be that as it may, this retrospective interpretation of item 3 is supported in Section 3.6.

My joint work with L. van der Torre on input/output logic was cumulative. Of all the papers we co-authored, item 4 represents the final key milestone we reached. The subsequent papers were further developments built on this foundation. The procedural semantics for out_3, incorporating a built-in consistency check and without closure of the output under logical consequence, is the most intricate of all. The discovery that the consistency check mirrors a consistency constraint limiting the application of the principle of (aggregative) cumulative transitivity was particularly striking. Proving the completeness theorem required a non-trivial extension of our previous proofs. In retrospect, we regret not having submitted the paper to a journal.

Item 5 presents an overview of my joint work with C. Benzmueller (and L. van der

Torre) on automated reasoning. Item 6 gives the latest developments, as well as a case study from population ethics, illustrating the relevance of my research to both computer science and philosophy.

1.5 List of (co-supervised) students

A great deal of my work would not have been possible without the input from the master and PhD students I co-supervised in Luxembourg and in Vienna. I served as the daily (co-)supervisor for all of them and also as the formal co-supervisor for D. Pichler.

Resulted in:

- 10 joint papers: [5, 8, 9, 10, 15, 32, 71, 72, 91, 100]
- 3 best paper awards: [5, 9, 100].
- 1 best master award (see below)

PhD

2024- Blaz Istenic Urh, TU Wien, Logic and Theory group. Topic: deontic explanations. Co-supervised with A. Ciabattoni.

2022- Dominik Pichler, TU Wien, Logic and Theory group. Topic: 1st-order deontic logic. Co-supervised with A. Ciabattoni.

2016-20 Ali Farjami, University of Luxembourg, Department of Computer Science. PhD title: *Discursive Input/output Logic*. Award date: 1 Oct 2020. Co-supervised with L. van der Torre.

2013-17 Diego Ambrossio, University of Luxembourg, Department of Computer Science. PhD title: *Non-Monotonic Logics for Access Control: Delegation Revocation and Distributed Policies*. Award date: May 2017. Co-supervised with L. van der Torre.

2012-16 Xin Sun, University of Luxembourg, Department of Computer Science. PhD title: *Logic and Games of Norms: a Computational Perspective*. Award date: July 2016. Co-supervised with L. van der Torre.

Master

2022 Dominik Pichler, TU Wien, Logic and Theory group. Master title: *Extensionality for Obligations in Åqvist's System F*. Co-supervised with A. Ciabattoni. Award date: 20 Oct 2022.

2019 Maya Olszewski, University of Luxembourg, Department of Computer Science. Master title: *Exploring Permission in I/O Logic*. Award date: 26 Aug 2019. Co-supervised with L. van der Torre.
Best master thesis in the Faculty of Science **

2018 Paul Meder, University of Luxembourg, Department of Computer Science. Master title: *Deontic Agency and Moral Luck*. Award date: 3 Sep 2018. Now high-school teacher. Co-supervised with L. van der Torre.

2016 Zohreh Baniasadi, University of Luxembourg, Department of Computer Science. Master title: *STIT Logic for Machine Ethics With IDP Specification*. Award date: 12 May 2016. Now in industry (name of the company: Zortify). Co-supervised with L. van der Torre.

The students I co-supervised had all received prior training in deontic logic through the Master's-level course I have taught for over ten years—initially in Luxembourg and subsequently in Vienna. This sustained teaching activity led to the publication of the textbook *Introduction to Deontic Logic and Normative Systems* [96], which now serves as a foundational resource for graduate students entering the field.

Chapter 2

Preference-based deontic logic

This section discusses my contributions to the axiomatization problem of Åqvist's preference-based dyadic deontic logics (DDLs).

The key idea is to analyze conditional ought (and more generally normative) statements in terms of a Lewis-type "intentional" conditional operator weaker than material (or even strict) implication as used in classical logic. Chisholm [26]'s well-known paradox of reparational or CTD obligation and akin have demonstrated, beyond any reasonable doubt, that normative conditionals as used in natural language behave differently from material implication. In propositional logic, the so-called paradoxes of material implication have long been taken to indicate that material implication cannot serve as a good formalization of if-then statements in actual usage. Consequently, there is a vast body of semantic and logical research focused on conditionals that differ from the material conditional. The same observation holds true for deontic logic and conditional norms. My PhD thesis [74] on logic and argumentation allowed me to thoroughly investigate this issue. The investigation carried out in there (and in the three companion papers [62, 73, 75]) shaped an understanding that has since become a fundamental assumption in my research.

2.1 Overview of the framework

For the purpose of this presentation, I confine myself to a brief description of the framework. The language has a conditional obligation operator $\bigcirc(-/-)$ and a S5-type alethic modal operator $\Box$ both viewed as a primitive. (In some systems among those studied here $\Box$ becomes definable in terms of $\bigcirc(-/-)$.) $\bigcirc(B/A)$ and $\Box A$ are

read as "If A, then B is obligatory" and "It is settled that A", respectively. The set of well-formed formulae (wffs) is defined in the usual way. There are no restrictions as to iterations of dyadic deontic operators and modal ones. $P(B/A)$ and $\Diamond A$ are a shorthand for $\neg \bigcirc (\neg B/A)$ and $\neg\Box\neg A$, respectively.

The semantics is in terms of so-called preference models. In models of this sort, a binary (preference) relation $\succeq$ ranks the possible worlds a, b, ... in terms of comparative goodness or betterness. For $a \succeq b$, read "a is at least as good as b". $\succ$ is the strict counterpart of $\succeq$, defined by $a \succ b$ ("a is strictly better than b") if $a \succeq b$ and $b \not\succeq a$. Let $\|A\|$ denote the truth-set of A, *i.e.*, the set of worlds at which A holds, and $\text{best}_{\succeq}(\|A\|)$ the subset of those that are best according to $\succeq$. The formulas $\Box A$ and $\bigcirc(B/A)$ are evaluated as follows, where M is a given preference model and a a world in it:

- $M, a \vDash \Box A$ iff $\forall b$ in M we have $M, b \vDash A$
- $M, a \vDash \bigcirc(B/A)$ iff $\forall b \in \text{best}(\|A\|)$ we have $M, b \vDash B$

Historically, this framework has strong connections with frameworks developed in related areas, like the theory of revealed preferences (as introduced by the economist Samuelson), the logic of conditionals (as developed in the 1970's following Stalnaker and Lewis) and the theories of nonmonotonic inference operations (as constructed in the 1980's in the context of logics for artificial intelligence). All these frameworks adopt a qualitative (or ordinal) approach. In the qualitative approach, possible worlds are ordered using a binary preference relation. Meaning (semantics) is defined through this preference relation, using maximality to select the most preferred options.

As mentioned in Chapter 1, Makinson [65] describes preference-based DDL as a "nonmonotonic formalism avant la lettre". Their first appearance in print was in Danielsson [34]'s PhD dissertation, published in 1968, and in Hansson [49]'s paper, which appeared in 1969. The different frameworks did not develop at the same pace, though. One can find a more detailed description of the framework in previous work. In particular, the reader may find it helpful to consult the chapter devoted to it in the *Handbook of Philosophical Logic* [3] and in the second volume of the *Handbook of Deontic Logic and Normative Systems* [83].

As mentioned, the main motivation for such a semantics has to do with the analysis of so-called contrary-to-duty (CTD) obligation sentences. They tell us what comes into force when some other (primary) obligations are violated. Since the discovery of Chisholm [26]'s CTD paradox and other paradoxes, a number of researchers in deontic logic have accepted the idea that an appropriate semantics for contrary-

to-duty obligation sentences calls for an ordering on possible worlds, in terms of preference or relative goodness.

Example 2.1.1 (Chisholm's paradox)**.** *Consider the following sentences: "You should go and help your neighbors"; "If you go, you should tell them you are coming"; "If you do not go, you should not tell them you are coming"; "You do not go". The first obligation is called a primary obligation, while the third one is called a contrary-to-duty obligation: it says what should be done if the primary obligation is violated. As is well-known, it is not possible to give a consistent formalization of these sentences in SDL. Fig 2.1.1 shows a preference-based model of these sentences, once formalized in DDL. The model tells us that go-and-tell is best, not-tell is second-best, and not-go-and-tell is worst. The formalization is consistent.*

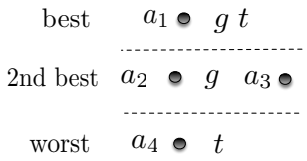

Figure 2.1.1: Chisholm's paradox. If a node is not labeled with an atomic sentence, then this one is false at this world

2.2 Gaps

I describe the two gaps my contributions have focused on.

Gap 1. The question of how to axiomatize dyadic deontic logic has been the focus of much attention, starting with van Fraassen [113] and Spohn [107], and continuing with Åqvist [3], Lewis [61] and Goble [44]. Existing results concerned classes of models whose relation $\succeq$ comes with the full panoply of the standard properties. These are: reflexivity, transitivity, totality (which rules out the possibility of incomparability) and different forms of the so-called limit assumption, ruling out sets of worlds without a best element (a "limit"). These proprieties (except reflexivity) have all been criticized as being too strong in some contexts. Totality is obvious. Several deontic logicians objected to the limit assumption, Lewis [61, pp. 97–98] among them. Transitivity was also criticized by economists and moral philosophers. There is a call for understanding what happens when these assumptions are not made.

Moreover, the properties of the betterness relation typically addressed in the literature do not represent the full range of available options. Additional candidate

properties deserve attention. In particular, various weakenings of transitivity–while extensively discussed in rational choice theory-have been almost entirely overlooked in deontic logic. These include:

- Quasi-transitivity. It demands that the strict part of the betterness relation be transitive: if $a \succ b$ and $b \succ c$, then $a \succ c$.
- Acyclicity: $\succ$ contains no cycles of the form $a_1 \succ a_2 \succ ... \succ a_n \succ a_1$.
- Suzumura consistency. This condition rules out cycles with at least one instance of strict betterness, e.g., $a_1 \succeq a_2 \succ a_3 \succeq ... \succeq a_n \succeq a_1$.
- Interval order. $\succeq$ is reflexive and Ferrers (if $a \succeq b$ and $c \succeq d$, then $a \succeq d$ or $c \succeq b$).

The interval order condition permits scenarios where transitivity of equal goodness fails, due to discrimination thresholds. These are cases where a and b are equally good, and b and c are equally good, but a and c are not equally good [63].

Gap 2. A number of researchers in so-called Rational Choice Theory have argued that the notion of "best" is ambiguous and that it can be characterized in three different ways: optimality, maximality and strong maximality. I will often refer to them as the opt rule, the max rule and the s-max rule, respectively. I have found that these different rules are not clearly distinguished in the deontic logic literature. For some world a to qualify as an optimal element of X, it must be as good as any other element in that set. For world a to count as a maximal element, no other world b in X must be strictly better than it. Thus, while the optimal elements are all equally good, the maximal elements are either equally good or incomparable. For a to be strongly maximal, no world b must be strictly better than any world c in a's equal goodness class. The max rule is often considered more suitable than the opt rule when incomparabilities between worlds is allowed. Due to Bradley [21], strong maximality is less widely known, but is arguably more suitable than maximality when the transitivity of betterness is no longer taken for granted. In particular, maximality leads to a violation of the requirement called "Indifference based choice" (IBC) by Bradley. It states that two options that are regarded indifferently (or equally good) should always be equally best. This is illustrated in Fig. 2.2.1, where b and c are equally good, but only c is maximal. Intuitively, it makes sense not to posit it as a best element. The problem is resolved by switching to strong maximality. Just as optimality and maximality collapse when $\succeq$ is assumed to be total, so also maximality and strong maximality collapse when $\succeq$ is assumed to be transitive.

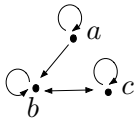

Figure 2.2.1: Violation of IBC. An arrow from a to b means that $a \succeq b$. No arrow from a to b means that $a \not\succeq b$.

In my [85], I give a concrete example, with a more complex structure. The pair $b \leftrightarrow c$ is replaced by a (finite) series of equally good worlds $b \leftrightarrow b_1 \leftrightarrow ... \leftrightarrow b_n \leftrightarrow c$. The example is different, because (like the relation of similarity) the equal goodness relation $\leftrightarrow$ is not assumed to be transitive. But the point being made is the same.

Table 2.2.1 summarizes the considered options:

max rule	**opt rule**	**s-max rule**
$\text{best}(\|\|A\|\|) = \max(\|\|A\|\|)$	$\text{best}(\|\|A\|\|) = \text{opt}(\|\|A\|\|)$	$\text{best}(\|\|A\|\|) = \max^s(\|\|A\|\|)$

Table 2.2.1: Notions of best

where

$$a \in \max(\|A\|) \Leftrightarrow a \models A \;\&\; \neg\exists b\,(b \vDash A \;\&\; b \succ a)$$
$$a \in \text{opt}(\|A\|) \Leftrightarrow a \models A \;\&\; \forall b\,(b \models A \rightarrow a \succeq b)$$
$$a \in \max^s(\|A\|) \Leftrightarrow a \models A \;\&\; \forall b\,\big((b \models A \;\&\; b \approx^A a) \rightarrow \neg\exists c\,(c \models A \;\&\; c \succ b)\big)$$

Here "$b \approx^A a$" means that b and a are two equally good A-worlds. But keep in mind that $\approx^A$ is only required to be reflexive.

Depending on what notion of best is used, one gets different evaluation rules for the conditional, but also different forms of the limit assumption.

As mentioned existing results in the literature [3, 61, 107, 113] were for classes of models whose relation $\succeq$ comes with the full panoply of the standard properties, in which case the different notions of "best" collapse. My task encompassed extending existing completeness results to broader classes of models. This involved conducting a systematic axiomatic study, taking into account the two key aspects mentioned above. Firstly, I explored variations in the properties of the betterness relation, covering the limiting case where no assumptions are made. Secondly, I examined the effects of using different notions of "best" in the truth-conditions for ought statements.

2.3 Results

At the semantic level, the options multiply exponentially. Perhaps the main finding is that the four systems in Fig. 3.2.1 can be used to bring order to the myriad of choices presented at the semantic level, hence providing a unified perspective.

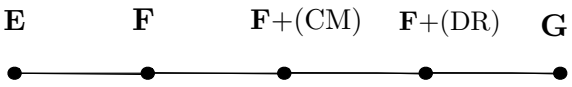

Figure 2.3.1: Systems

The systems are of increasing strength. A line between two systems indicates that the system to the left is contained in the system to the right. The base system is Åqvist's system **E** [2]:

All truth functional tautologies	(PL)
S5-schemata for $\Box$ and $\Diamond$	(S5)
$\bigcirc(B \to C/A) \to (\bigcirc(B/A) \to \bigcirc(C/A))$	(COK)
$\bigcirc(B/A) \to \Box \bigcirc(B/A)$	(Abs)
$\Box A \to \bigcirc(A/B)$	(Nec)
$\Box(A \leftrightarrow B) \to (\bigcirc(C/A) \leftrightarrow \bigcirc(C/B))$	(Ext)
$\bigcirc(A/A)$	(Id)
$\bigcirc(C/A \wedge B) \to \bigcirc(B \to C/A)$	(Sh)
If $\vdash A$ and $\vdash A \to B$ then $\vdash B$	(MP)
If $\vdash A$ then $\vdash \Box A$	(N)

(COK) is the conditional analogue of the familiar distribution axiom K. (Abs) is the absoluteness axiom of Lewis [61], and reflects the fact that the ranking is not world-relative. (Nec) is the deontic counterpart of the familiar necessitation rule. (Ext) permits the replacement of equivalent sentences in the antecedent of deontic conditionals. (Id) and (Sh) are familiar from the literature on non-monotonic logic. (Id) is the deontic analogue of the identity principle. The question of whether this is a reasonable law for deontic conditionals has been much debated. A defense of (Id) can be found in Hansson [49] This line of defense is discussed in Parent [79]. (Sh) is named after Shoham [105], who seems to have been the first to discuss it. It corresponds to the so-called "conditionalization" principle (also referred to as "the

hard half of the deduction theorem"), which is part of Kraus and colleagues' system C for cumulative inference relations (see [57]).

Next comes Åqvist's system **F**, obtained by supplementing **E** with the law (D*): $\Diamond A \rightarrow (\bigcirc(B/A) \rightarrow P(B/A))$, where (D*) is the dyadic analog of the D axiom. $\Diamond$ is the dual of $\Box$, and P is the dual of $\bigcirc$. Then comes **F**+(CM), obtained by supplementing **F** with the principle of cautious monotony (CM): $(\bigcirc(B/A) \wedge \bigcirc(C/A)) \rightarrow \bigcirc(C/A \wedge B)$. (CM) is well-known from the non-monotonic logic literature [57]. Intuitively (CM) says that complying with an obligation does not modify our other obligations arising in the same context. Next, we have **F**+(DR), obtained by supplementing **F** with the principle of disjunctive rationality [57]: $\bigcirc(C/A \vee B) \rightarrow (\bigcirc(C/A) \vee \bigcirc(C/B))$. (DR) tells us that if a disjunction of state of affairs triggers an obligation, then at least one disjunct triggers this obligation. Last, we have Åqvist's system **G**; it is obtained by supplementing **F** with the principle of rational monotony (RM) [60]: $(P(B/A) \wedge \bigcirc(C/A)) \rightarrow \bigcirc(C/A \wedge B)$. (RM) tells us that realizing a permission does not modify our other obligations arising in the same context.

F+(CM) is to **G** what the KLM system P is to their system R [57, 60]. **F**+(DR) is the conditional logic analog of Freund's disjunctive consequence relations [36, 20]. Note the so-called principle of consistency preservation—the analog of (D*), with consistency used as a surrogate of possibility—is not part of the latter three systems.

I present an overview of the key findings through two tables.

Table 2.3.1 deals with the distinction between maximality and optimality, whereas Table 2.3.2 focuses on the distinction between maximality and strong maximality, along with the different weakenings of transitivity mentioned above. In both tables, the leftmost column shows the property (or properties) of the betterness relation, and the other two columns show the corresponding systems under the relevant rule of interpretation. Limitedness and smoothness are the two forms of the limit assumption previously mentioned. Here is their exact formulation.

Limitedness If $\exists a.\ a \models A$, then $\text{best}_{\succeq}(\|A\|) \neq \emptyset$;

Smoothness If $a \models A$ and $a \notin \text{best}_{\succeq}(\|A\|)$, then $\exists b \succ a$ with $b \models A$ and $b \in \text{best}_{\succeq}(\|A\|)$.

It is understood that in each column limitedness and smoothness are cast in terms of the appropriate notion of best.

Property of $\succeq$	max	opt
–		
+ reflexivity + totality	**E**	**E**
transitivity + reflexivity	**E**	–
limitedness + reflexivity + totality	**F**	**F**
smoothness + reflexivity + totality	**F**+(CM)	**F**+(CM)
smoothness + transitivity ± reflexivity	**F**+(CM)	**G**
+ transitivity and totality	**G**	

Table 2.3.1: Opt vs max

Each cell corresponds to a strong soundness and completeness theorem. I recall that strong completeness establishes a match between the syntactical and semantical consequence relation while accommodating a potentially infinite set of assumptions. For instance, the top row in Table 2.3.1 expresses the following theorems:

Theorem 2.3.1: Completeness of E

Under the opt rule and the max rule, **E** is strongly sound and complete with respect to the following three classes of preference models:

1. the class of all preference models;
2. the class of those whose relation $\succeq$ is reflexive;
3. the class of those whose relation $\succeq$ is total.

The reading of the other rows is similar.

One can see that the contrast between maximality and optimality is not as significant as one would expect, because in most cases the determined logic remains the same. Under the max rule or the opt rule, **E** is sound and complete with respect to the class of all preference models, the class of those in which $\succeq$ is required to be reflexive, and the class of those in which it is required to be total. Second, under either rule, **F** is sound and complete with respect to the class of preference models in which $\succeq$

is required to be limited, the class of those in which it is required to be limited and reflexive, and the class of those in which it is required to be limited and total. Third, under the max rule or the opt rule, **F**+(CM) is sound and complete with respect to the class of models in which $\succeq$ is required to be smooth, the class of those in which it is required to be smooth and reflexive, and the class of those in which it is required to be smooth and total. Likewise, under either rule, **G** is sound and complete with respect to the class of models in which $\preceq$ is required to be smooth, transitive and total (and hence, reflexive).

The observation that the contrast between maximality and optimality is less significant than one might expect requires qualification. There are two cases where the choice between these two notions of "best" makes a difference. Both cases have to do with the transitivity of the betterness relation:

- In the presence of smoothness, it appears that, under the opt rule, transitivity alone lifts the logic from **F**+(CM) to **G**, whereas under the max rule, both transitivity and totality are needed to get **G**.
- In the absence of smoothness, and even limitedness, it appears that, under the max rule, transitivity alone does not affect the logic, while under the opt rule transitivity boosts the logic from **E** to **E**+(RM)+(transit), where (transit) expresses a principle of transitivity for a weak preference relation $\geq$ over formulas defined by $A \geq B := P(A/A \vee B)$:

$$P(A/A \vee B) \wedge P(B/B \vee C) \rightarrow P(A/A \vee C) \qquad \text{(transit)}$$

  (Result initially established by Grossi & al. [45]).

Now, I move to Table 2.3.2. This one shows that, to some extent, the point made about "optimality vs maximality" carries over to "maximality vs strong maximality". First, it appears that, regardless of the presence or absence of any form of the limit assumption, the choice between maximality and strong maximality does not affect the logic—except in one case. Additionally, quasi-transitivity, acyclicity, Suzumura consistency, and full transitivity appear to have no impact on the logic, independently of the limit assumption. However, the distinction between maximality and strong maximality becomes significant when considering the interval condition. Under the max rule, the interval condition boosts the logic from **F**+(CM) to **F**+(DR), while under the *s*-max rule, it boosts the logic from **F**+(CM) to **G**.

The exact nature of the result established for **F**+(DR) may not be immediately evident; therefore, I will state it explicitly. This is Theorem 2.3.2. A model is

Property of $\succeq$	max	s-max
$-$		
+ acyclicity		
+ Suzumura consistency	**E**	**E**
+ quasi-transitivity		
+ transitivity		
limitedness		
+ acyclicity		
+ Suzumura consistency	**F**	**F**
+ quasi-transitivity		
+ transitivity		
smoothness		
+ acyclicity		
+ Suzumura consistency	**F**+(CM)	**F**+(CM)
+ quasi-transitivity		
+ transitivity		
interval order[a]	**F**+(DR)[b]	**G**[c]

Table 2.3.2: Max vs s-max

[a] With limitedness
[b] Weak completeness w.r.t. finite models and weak completeness *tout court*
[c] Weak completeness w.r.t. finite models and strong completeness

said to be finite if it has finitely many worlds. Weak completeness establishes a correspondence between validities and theorems.

Theorem 2.3.2: Interval order

Under the max rule, **F**+(DR) is weakly complete with respect to the following two classes of models:

1. The class of finite models whose relation $\succeq$ is an interval order;
2. The class of models whose relation $\succeq$ is an interval order and is limited.

The result also holds under the opt rule. Strong completeness under the max rule is open. By contrast, the result established under the s-max rule is both a weak completeness one over the class of finite models, and a strong completeness result.

Decidability of the theoremhood problem ("Is A a theorem?") has been established

along the way for all the systems.

Publications: [80, 76, 81, 83, 84, 85]

2.4 Advancements beyond deontic logic

These results improve the state-of-the-art in dyadic deontic logic, but also in the related areas of nonmonotonic logic and the logic of counterfactuals.

First, they complement those of Makinson, and Kraus, Lehmann & Magidor (KLM) for non-monotonic inference relations [57, 60, 64, 66]. These authors assume the form of the limit assumption called smoothness (or stopperedness). As mentioned such an assumption has been criticized, notably by Lewis [61]. There is a call for understanding what happens in its absence. On the other hand, these authors work with a primitive strict or irreflexive relation $\succ$ ("strictly better than") in the models while I use a primitive non-strict relation $\succeq$ ("at least as good as") as is the custom in deontic logic. The advantage of using a primitive non-strict relation is that one can more easily distinguish between worlds that are tied or equally good and worlds that are incomparable. (See the discussion in [85].) One can then provide a finer-grained semantical analysis, and disentangle different notions of "best" like the above three, and different candidate weakenings of transitivity. Only maximality and quasi-transitivity are considered by these authors.

The same applies to the study of the interval order condition recently conducted by Booth and Varzinczak [20] within the KLM setting.

My completeness proofs for **F**+(CM) and **G** have drawn on the one for system P in [57]. The axiom (CM) plays an essential role. The completeness proof for **E** and **F** required new methods. I have used the two-step methodology initially developed by Schlechta [104]. One makes a detour through a semantics in terms of a selection function f, assigning to each set of worlds its "best" elements [25]. The first step consists in establishing completeness with respect to classes of models equipped with a selection function. The second step consists in showing that the selection function semantics matches the preference semantics. Here the main difficulty is to show how to derive the latter from the former. This amounts to showing that, starting with a selection function model of the appropriate kind, one can always generate an equivalent preference model of the appropriate kind. This is not straightforward. The obtained results generalize similar ones established in rational choice for the particular case where the axioms of rational monotony (RM) and (D*) are available.

This method has also led to a completeness proof for **F**+(DR) that differs from the one given by Booth and Varzinczak [20] for disjunctive consequence relations.

My findings also offer a complementary perspective to those presented by Lewis [61] in his work on the logic of counterfactuals. Lewis' prime interest is in strong logics of counterfactuals, for which the similarity relation is transitive and total. Thus he does not address models that might be appropriate for weaker systems. In [61] Lewis directs our attention to the systems that might be suitable for a logic of conditional obligation. He designates his own preferred deontic system as **VN**, providing it with various modelings, with the "preferred" one being formulated in terms of sphere models. Transitivity is inherently embedded within a system of spheres, corresponding to the nesting property among these spheres. It remains an interesting question whether our proposed relaxations of transitivity find analogous counterparts within this framework. In his [61], he proposes two variant evaluation rules for the conditional in terms of betterness, allowing to avoid some of the side effects of letting the limit assumption go. It is not known what happens when transitivity is relaxed.

2.5 Postcript on (hyper-)sequent calculi and complexity

Gentzen-style sequent calculi are known to be better suited for automated reasoning. With A. Ciabattoni and N. Olivetti, we have investigated **E** and **F** from a sequent calculus viewpoint. In [27, 28], we have proposed a hypersequent calculus for **E** and **F**, with the cut-elimination property and the subformula property. Hypersequents constitute a natural generalization of ordinary sequents. They may be thought as finite multisets (or sequences) of usual sequents. They are needed to prove cut-elimination for S5, a subsystem of **E** and **F**. The obtained property is cut-admissibility: if a hyper-sequent is derivable using the cut rule, then it is also derivable without it.

For **E**, the calculus has been refined to obtain a complexity result, stating that the validity problem in **E** ("Is the formula A valid?") is Co-NP, like in classical propositional logic. Therefore, conditional normative reasoning is not harder than ordinary (propositional) reasoning although it requires a more expressive language. This fact was not known before. It makes an essential use of the fact that the betterness ranking is not world-dependent, so all worlds agree on all conditional statements. The result echoes one previously obtained by Friedman and Halpern

[37] in the related area of conditional logic. However, they work with models whose preference relation is a pre-order (is reflexive and transitive), and they use a more complex evaluation rule for the conditional, of the form $\exists\forall\exists$. It puts $\bigcirc(B/A)$ as true, whenever for all A-worlds there is an A-world above it starting a (possibly infinite) sequence of increasingly better worlds where $A \rightarrow B$ holds. The follow-up paper [28] reports a first complexity bound for system **F**. Deciding if a formula is a theorem of **F** is CoNEXP. An efficient method for extracting a counter-model from a failed proof attempt has been discovered for **E**, but not for **F**. The problem is with the limitedness condition; this one is not a frame condition.

One notable feature of the framework is its use of a unary $\mathcal{B}et$ modality. Like in previous work on modal interpretation of conditionals, *e.g.*, [41], we use this modal operator to encode maximality in the syntax. Intuitively the fact that a is among the best A-worlds may be understood as saying that all the worlds accessible from a via the betterness relation (or "above" a according to the ranking) falsifies A. This is encoded as $\mathcal{B}et\neg A$, where $\mathcal{B}et$ is a K-type modal operator. The conditional obligation $\bigcirc(B/A)$ can then be indirectly defined as $\Box(A \wedge \mathcal{B}et\neg A \rightarrow B)$, where $\Box$ obeys the laws of S5.

The $\mathcal{B}et$ modality is needed formulate appropriate rules of introduction and elimination of $\bigcirc(-/-)$. Below I give the example of the right introduction rule of $\bigcirc(-/-)$, here cast in an ordinary sequent calculus:

$$\frac{\Gamma^{\Box}, \Gamma^{O}, A, \mathcal{B}et\neg A \Rightarrow B}{\Gamma \Rightarrow \bigcirc(B/A), \Delta}\ (\bigcirc\ \mathrm{R})$$

where

$$\Sigma^{O} = \{\bigcirc(C/D) : \bigcirc(C/D) \in \Sigma\} \text{ and } \Sigma^{\Box} = \{\Box A : \Box A \in \Sigma\}.$$

A sequent $\Gamma \Rightarrow \Delta$ corresponds to the validity of the formula $\Box(\wedge\Gamma \rightarrow \vee\Delta)$ in the preference model. The rule allows to introduce $\bigcirc(B/A)$ at the right hand side of $\Rightarrow$ if $\Box(A \wedge \mathcal{B}et\neg A \rightarrow B)$ has been established.

Publications: [27, 28]

2.6 Application

See Section 4.4.

Chapter 3

Norm-based deontic logic: I/O logic

I describe my contributions to I/O (input/output) logic, which falls within the category of what Hansen [48] has called a "norm-based" deontic logic. The core idea of a norm-based deontic logic is to explain the truths of deontic logic not by some set of possible worlds among which some are ideal or at least as good as others, but with reference to an explicit set of given norms or existing moral standards.

As already mentioned, a conditional obligation is represented as a pair (a, x) of Boolean formulae, where a and x are the body (antecedent) and the head (consequent), respectively. A normative system N is a set of such pairs. The main semantical construct is given the following general form: $x \in out(N, A)$, where A is a set of formulas. This is read as follows: given input A (state of affairs), x (obligation or permission) is in the output under norms N. The central question is: what obligations (resp. permissions) can be detached from a set N of (explicitly given) rules or conditional norms in a given context? The approach is in this regard different from the more traditional one, aiming at identifying a set of "logical laws" using a possible worlds semantics.

The purpose of the semantics is to define the function *out*. It does so through a set of procedures that generate outputs based on given inputs. The semantic framework relies on the concept of detachment, and is inherently complex. When available, the proof theory is relatively simple, and more user-friendly. It operates through a set rules manipulating pairs of formulas (a, x) rather than individual formulas. But there is more to I/O logic than just deriving pairs from pairs. When a derivation

contains a node of the form (a, x), under the hood the reasoner has calculated that $x \in out(N, a)$.

This chapter is organized as follows. Section 3.1 explains the main idea underpinning the semantics. Section 3.2 gives some background on the traditional framework. Section 3.3 summarizes my contributions. Sections 3.4, 3.5 and 3.6 describe them in more details.

3.1 A non-committal semantics

Makinson was motivated by the desire to reconstruct deontic logic in a way that reflects the view that norms do not possess truth-values. While this was a primary motivation for him, I will not place as much emphasis on it.

Whether norms admit truth-values, and whether a possible-world semantics inevitably entails affirming that they do, are both highly intricate matters—and I will not be addressing them here.

Rather than focusing on the truth-status of norms, I prefer to highlight a different and, in my view, more fundamental feature of the input/output (I/O) framework: the central role of modus ponens or factual detachment in its semantics. Modus ponens is familiar from propositional logic. It states that from a conditional statement and its antecedent, one may infer the consequent. This is the rule, where $\rightarrow$ denotes material implication:

$$\frac{a \qquad a \rightarrow b}{b} \text{ (Modus ponens)}$$

In deontic logic, this is often referred to as factual detachment, because the antecedent whose truth is required expresses a fact.

The central role of detachment is also present in Horty [53]'s deontic default logic. I/O logic shares this feature with it. The bottom line between the two frameworks is that I/O logic does not allow throughput, viz. it does not allow the input to be in the output. Proof-theoretically, this means that I/O logic does not contain the rule:

$$\frac{-}{(a, a)} \text{ (ID)}$$

The preference-based approach to conditional obligation described in Chapter 2 shares with the so-called classical theory of rational choice [103] the assumption

that an individual has (well-defined) preferences, and that a normative judgment is based on a maximization process. Advocates of (as Simon [106] terms it) "bounded rationality" have argued that an approach to rationality in terms of maximization is not realistic, because human beings lack the cognitive resources to optimize. Second, such a logic takes the so-called trichotomy thesis for granted. It is the assumption that comparable items (worlds) can only be better than, worse than, or equal to each other in overall value. The trichotomy view has been challenged more than once. Some have argued that these three value relations do not exhaust the space of possibilities. The most well-known proposal for a fourth *sui generis* relation is Chang's argument for "on a par"−see Chang [24].

The best way to avoid potential objections is to make as few assumptions as possible. The assumption that conditionals obey the rule of modus ponens or detachment is one that can hardly be challenged. Obligations and permissions are contextual and vary based on the setting. Consequently, a norm always takes the form of a conditional statement. Some philosophers like Boghossian [19] think (rightly, in my view) that the disposition to reason according to detachment is constitutive of the possession of the concept of conditional, and thus of the concept of norm. The idea is that, if some agent says "if a then x", and if s.he truly means it, then s.he commits her.himself to detaching x given a. If this agent refuses to acknowledge that s.he is justified in employing detachment, this will be good evidence that s.he fails to understand what is meant by "if ... then". Accepting detachment and acquiring an implication are simply two sides of the same coin.

Publications: [92, 94]

3.2 Background on the classical framework

This section gives some background on the classical framework as initially devised by Makinson and van der Torre [67, 68, 69]. It has two parts or levels, called "unconstrained" and "constrained" I/O logic.

A rough analogy with the dyadic deontic logics discussed in Chapter 2 can be expressed as follows (omitting curly brackets since this is a singleton input set):

$$x \in out(N, a) \leftrightarrow N \models \bigcirc(x/a)$$

While this is merely an analogy, I have recently discovered a way to turn it into a mathematical result. (See Chapter 5.)

Name	Semantics	Rules
out_1 (simple-minded)	1-step detachment	SI,WO,AND
out_2 (basic)	disjunctive input	SI,WO,AND,OR
out_3 (reusable)	iterated detachment	SI,WO,AND,CT
out_4 (basic reusable)	out_2+out_3	SI,WO,AND,OR,CT

Table 3.2.1: Traditional I/O operations

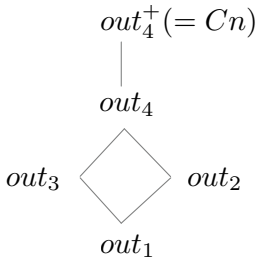

Figure 3.2.1: Relations

3.2.1 Unconstrained I/O logic

Four I/O operations, based on the notion of detachment, have been proposed in [67, 69], yielding to four systems of increasing strength. They are shown in Table 3.2.1 and Figure 3.2.1. The associated proof systems are displayed too. Their rules are shown in Table 3.2.2, where $\vdash$ denotes the relation of consequence in classical propositional logic. In Figure 3.2.1, the fifth I/O operation out_4^+ makes the connection with classical logic, more specifically the operation Cn defined by setting $Cn(A) = \{x : A \vdash x\}$, where A is a set of formulas.

out_1 is the weakest I/O operation, based on a 1-step detachment. One takes the heads of all the pairs whose body is classically implied by the input, and closes the set under logical consequence. out_2 extends out_1 to handle a disjunctive input formula, of the form $a \vee b$. It is associated with reasoning by cases on the input, also known as disjunctive reasoning or proof by cases. One considers in turn each maximal consistent extension of $a \vee b$, applies out_1 to each input set, and takes the meet of the resulting output sets. out_3 extends out_1 to iteration of successive detachments. This allows chaining of norms and output reuse. out_4 combines out_2 and out_3.

The proof theory is given in terms of inference rules manipulating pairs of the form (a, x), like in conditional logic. $x \in out(N, a)$ is assumed to be equivalent with $(a, x) \in out(N)$. This "trick" enables the formulation of a proof system.

$\dfrac{(a,x) \quad (a,y)}{(a, x \wedge y)}$ AND	$\dfrac{(a,x) \quad (b,x)}{(a \vee b, x)}$ OR
$\dfrac{(a,x) \quad (a \wedge x, y)}{(a,y)}$ CT	$\dfrac{(a,x) \quad b \vdash a}{(b,x)}$ SI
$\dfrac{(a,x) \quad x \vdash y}{(a,y)}$ WO	$\dfrac{-}{(a,a)}$ ID

Table 3.2.2: Rules. The labels "SI", "WO" and "CT" stand for "Strengthening of the input", "Weakening of the output" and "Cumulative Transitivity" respectively.

Soundness and completeness results establish the equivalence between the syntactical and semantical characterizations.

Each I/O operation out_i has a throughout version out_i^+, in which the input reappears as output. Semantically, one looks at what is outputted by $N \cup \{(a,a) : a \text{ is a formula}\}$. Syntactically, one adds the rule

$$\frac{-}{(a,a)} \text{ (ID)}$$

out_4^+ (called throughout basic reusable) collapses with classical logic, in the following sense: $x \in out_4^+(N, A)$ iff $x \in Cn(A \cup m(N))$, where $m(N)$ is the materialization of N, viz $\{a \to x : (a,x) \in N\}$.

At least two concepts of permission can be distinguished: weak (or negative) permission, and strong (or positive) permission. x is weakly permitted in context a, if it is not forbidden. Strong permission requires the introduction of a set P of explicit permissions. x is strongly permitted in context a, if x is outputted under N and P given input a. The proof theory of permission is obtained from the proof theory of obligation by constructing suitable rules that correspond to the subverse and inverse of those governing obligation.

3.2.2 Constrained I/O logic

Unconstrained I/O logic faces challenges such as Chisholm's paradox, which have led to the development of constrained versions of I/O logic [68].

Example 3.2.1 (Chisholm's paradox)**.** *Put* $N = \{(\top, g), (g, t), (\neg g, \neg t)\}$*, where* g *and* t *denote the proposition that "John goes and helps his neighbors" and "John tells them that he is coming". When the chaining of norms is allowed (*out_3 *and* out_4*), under input* $\neg g$*, the contradictory outputs (*t *and* $\neg t$*) are obtained, and so any state of affairs (e.g., kill the boss) becomes obligatory.*

The basic idea underpinning constrained I/O logic is familiar from belief change theory. One makes the minimal changes to the normative system N to restore consistency. In other words, one cuts down N to just "below" the threshold of yielding excess, and then restrict the output accordingly. This requires the use of a set C of constraints (= formulas), hence the name "constrained I/O logic". The output is required to be consistent with it. This is implemented using the notions of "maxfamily" and "outfamily".

- maxfamily: $\mathit{maxf}(N, A, C)$ is the set of $\subseteq$-maximal subsets H of N such that $\mathit{out}_i(N, A)$ is consistent with C.
- outfamily: $\mathit{outf}(N, A, C) = \{\mathit{out}_i(H, A) \mid H \in \mathit{maxf}(N, A, C)\}$.

The full meet constrained output under N given input A (= our final output) and constraint C is defined as $\cap \mathit{outf}(N, A, C)$ (skeptical approach) or as $\cup \mathit{outf}(N, A, C)$ (credulous approach). In CTD scenarios, the outfamily has one element only, so the skeptical and credulous approach yield the same output. Also C is assumed to be A. The input is assumed to be settled as true, and cannot be changed. Therefore the output is required to be consistent with it ("ought" implies "can").

Example 3.2.2 (Chisholm's paradox, cont'd)**.** *Let* N *be as in Example 3.2.1. The maxfamily has* $\{(g, t), (\neg g, \neg t)\}$ *as its sole element. This is the biggest subset of* N *whose output is consistent with* $\neg g$*. The full meet constrained output (=the final output) is* $Cn(\neg t)$*.*

3.3 Summary of contributions

The main contribution of my research includes:

(Unconstrained level) A systematic axiomatic study of (unconstrained) I/O logic systems derived from the chosen input/output operations for obligation and permission. A final tangible output is the I/O analog of the modal cube.

(Constrained level) An extension of constrained I/O logic with priorities, a proof system for some I/O operations, and a faithful embedding of preference-based dyadic deontic logic within constrained I/O logic.

(Both) An application to Kelsen's conception of norms and to moral particularism (following Horty's footsteps).

In the following sections I describe these contributions in more details.

3.4 Systematic axiomatic study

The goal of this work is similar to the goal pursued in the study of preference-based dyadic deontic logic: to carry out a systematic axiomatic study of I/O logic systems, focusing on variations arising from different choices of I/O operations for obligation and permission. This necessitated extending and generalizing the classical framework described in Section 3.2.

This is joint work with L. van der Torre, and in part with D. Gabbay and M. Olszewski, a master's student. The most challenging aspects involved formulating suitable semantic definitions, identifying corresponding axiomatic systems, and extending the original completeness proof to accommodate the new framework.

Four new families of systems have been defined. Each family is presented in the order it was developed.

3.4.1 Intuitionistic I/O logics. The base logic of I/O logic is classical propositional logic. Independent arguments can be supplied to support the idea that intuitionistic logic is more suitable than classical logic in the legal domain. In particular, a case can be dismissed due to insufficient evidence, indicating that the principle of the excluded middle ("$A \vee \neg A$") does not hold.

In collaboration with D. Gabbay and L. van der Torre, we have investigated what happens when the base logic is changed from classical to intuitionistic logic. The main finding is that, when going intuitionistic, a representation theorem is still available for the first three (unconstrained) original I/O operations. The primary challenge arises with out_2, whose semantics relies on the notion of maximal consistent set, which is no longer available in intuitionistic logic. The key idea is to replace the notion of maximal consistent set by its intuitionistic analog, the notion of saturated set. Unlike a maximal consistent set, a saturated set does not necessarily contain either a formula or its negation. The axiomatic characterization is as in the classical case (up to the logical connectives). Therefore, the choice between the two base logics does not make any difference for the resulting framework. In a sequel, I have

shown that intuitionistic I/O operation out_2 can faithfully be embedded in the so-called constructive modal logic CK devised by de Paiva and colleagues. Roughly, one shows that x is outputted by N under input A if and only if, in CK, $\{b \to \Box y : (b, y) \in N\} \cup A$ proves $\Box x$, where $\to$ is intuitionistic implication and $\Box$ is a K-type modality.

Publications: [82, 90].

3.4.2 Systems without output closure (WO). The traditional I/O operations validate the rule WO, stating that, if a state of affairs is obligatory, then so are its logical consequences.

$$\frac{(a, x) \qquad x \vdash y}{(a, y)}\ \text{WO}$$

Like in epistemic logic, in deontic logic the intuitive desirability of such a law has been called into question. This has led us to redefine the four traditional I/O operations in such a way that they no longer satisfy this principle.

On the proof theoretical side, the basic idea is to replace WO with the following rule, which expresses a principle of replacement of logically equivalents for the output ($x \dashv\vdash y$ is short for $x \vdash y$ and $y \vdash x$):

$$\frac{(a, x) \qquad x \dashv\vdash y}{(a, y)}\ \text{EQ}$$

On the semantical side, one requires the output x to be logically equivalent with the conjunction of the rules in N used to get x. The main challenge has been to work out a suitable definition for out_2 and out_4, and establish soundness and completeness of the systems with respect to the semantics.

I briefly motivate the removal of WO. The rule yields as a special case the principle of conjunction elimination, warranting the move from $(a, x \wedge y)$ to (a, x). As suggested for example by Hamblin [46], Goble [43] and Hansen [48], such a principle is counter-intuitive in those cases where x and y are not separable, so that "failing a part means that satisfying the remainder no longer makes sense" [48, p. 292].

Example 3.4.1 (Shopping list, [48])**.** *Suppose I have the obligation to buy apples and walnuts, these being meant to land in a Walforf salad. It might be unwanted and a waste of money to buy the walnuts if I cannot get the apples. Thus, the obligation to buy apples and walnuts does not imply the obligation to buy the walnuts.*

3.4.3 Systems with an aggregative form of cumulative transitivity (ACT). We have also studied a family of variant systems with a new form of transitivity between norms. They make sense only in a system without WO. Therefore, there is an overlap with the second family of systems.

The traditional systems corresponding to out_3 and out_4 contain the rule called "Cumulative Transitivity" (CT). This is the traditional form of transitivity considered in the literature. It implies the principle called "Deontic Detachment" (DD):

$$\frac{(a,x) \qquad (a \wedge x, y)}{(a,y)}\ \text{CT} \qquad\qquad \frac{(\top,x) \qquad (x,y)}{(\top,y)}\ \text{DD}$$

Counter-examples have been given to DD (see, e.g. [22, 58], and below). Such counter-examples appeal to the seemingly plausible idea that, in context a (resp. $\top$), the obligation of y does not hold, if the obligation of x is violated. Therefore, they can be blocked by replacing CT with the rule "Aggregative Cumulative Transitivity (ACT)", which keeps track of previously detached obligations.

$$\frac{(a,x) \qquad (a \wedge x, y)}{(a, x \wedge y)}\ \text{ACT}$$

As mentioned, this substitute rule makes sense only in a system without WO. This is because, given WO, ACT implies CT. Here is an example showing the plausibility of ACT.

Example 3.4.2 (The pay and display machine, [71])**.** *The Luxembourgish traffic laws [1] say that if one wants to park one's car at a parking spot having a park meter during the times specified on the street sign, then one should buy a ticket. They also say that, if a parking ticket is purchased, then it should be put on display inside the vehicle. The machine selling the tickets is usually called a "pay and display machine". The obligation to put the ticket on display no longer holds, if the obligation to pay is violated (for instance the ticket has been forged). Thus, the correct conclusion is: one should pay-and-display the ticket.*

Publications: [93, 94, 98].

3.4.4 Systems with a built-in consistency check. We have defined and studied I/O logics whose I/O operation has a built-in consistency check. Their syntactical counterparts are a restricted version of AND and ACT below, R-AND and R-ACT.

$$\frac{(a,x) \qquad (a,y) \qquad a,x,y \not\vdash \bot}{(a, x \wedge y)}\ \text{R-AND} \qquad \frac{(a,x) \qquad (a \wedge x, y) \qquad a,x,y \not\vdash \bot}{(a, x \wedge y)}\ \text{R-ACT}$$

Note that, in the presence of SI, R-ACT implies R-AND.

The motivation for R-AND is the need to eliminate the pragmatic oddity without creating the drowning problem.

Example 3.4.3 (Drowning problem, [101])**.** *The drowning problem arises when a primary obligation no longer holds after a violation has occurred. This is usually considered problematic. Consider the unconditional obligation to keep a promise, $N = \{(\top, k)\}$. This obligation still holds when violated. One should have $k \in out(N, \neg k)$. On the proof theoretical side, to avoid the drowning problem requires keeping the rule "Strengthening of the input" (SI).*

$$\frac{(\top, k)}{(\neg k, k)}\ SI$$

The pragmatic oddity then arises if an unrestricted form of aggregation is allowed.

Example 3.4.4 (Pragmatic oddity, [101])**.** *Consider the same N as in Example 3.4.3, but add the extra obligation "if you do not keep your promise, then you should apologize", $(\neg k, a)$. Using AND, one derives the (rather problematic) unconditional obligation to keep your promise and apologize for not keeping it.*

$$\frac{\dfrac{(\top, k)}{(\neg k, k)}\ SI \qquad (\neg k, a)}{(\neg k, k \wedge a)}\ AND$$

The inference may be blocked using R-AND, and by not allowing the aggregation of the two obligations. Intuitively, this is justified, because $(\neg k, k)$ and $(\neg k, a)$ do not express the same kind of obligation. $(\neg k, a)$ expresses what has been called an actual obligation (what is obligatory, given what is actually the case), while $(\neg k, k)$ expresses an ideal obligation (what should have been done).

The motivation for R-ACT is to allow a form of transitivity between norms, while at the same time restricting aggregation the way just described. The combination (R-AND, CT) is a priori possible, but we found it difficult to define a suitable I/O operation satisfying these two rules.

Intuitively, the semantics captures the idea of (as we call it) "backtracking". To determine if x is in the output, one goes back in time, before the violation has

occurred, and check what norm applied at that time. This is why only the move to $(\neg k, k)$ is warranted. Before the violation has occurred, viz. in context $\top$, the obligation of k applied, but not the obligation of $k \wedge a$.

The consistency check is different from the one in constrained I/O. Say a rule (a_i, x_i) is directly grounded in input a if a logically entails a_i. The difference is this: instead of requiring the output to be consistent with the input (and hence consistent per se), we require the output to be consistent with the bodies of the rules that are directly "grounded" in the input.

Backtracking: reusable case (output reused as input)

x is outputted given input a if and only if

1. x is logically equivalent with the conjunction of the heads of finitely many pairs (a_1, x_1), ..., (a_n, x_n) grounded directly or indirectly (modulo chaining) in a, and
2. x is consistent with the set of all the bodies of those pairs that are directly grounded in a.

Example 3.4.5 (Backtracking)**.** *In the previous two examples, given input $\neg k$, k is outputted. This output is consistent with the body of the rule used to obtain it, namely $(\top, k)$. This is how the drowning problem is avoided. But $(\top, k \wedge a)$ is not outputted, because $k \wedge a$ is not consistent with the body of the rule $(\neg k, a)$. This is how the pragmatic oddity is blocked.*

A soundness and completeness theorem shows the equivalence between the semantics and the proof system. Among all the completeness results we obtained, this one was the most difficult to obtain.

With Maya Olszewski, we have extended the account to (strong) permission, and obtained a series of characterization results, based on establishing the so-called non-repetition property. It says that, if (a, x) is derivable from N, then it is derivable from N using a member of N at most once.

The work on permission was based on Maya's master, which we co-supervised, and which was awarded the best master thesis award within the faculty.

Publications: [71, 72, 95, 97].

3.5 Priorities, proof systems, link with preferences

3.5.1 Prioritized I/O logic. I have extended constrained I/O logic to support reasoning about prioritized obligations. Reasoning about prioritized obligations refers to the process of reasoning about situations where multiple duties exist, but some are more important than others.

The obtained framework is meant as an alternative to Horty [54]'s default deontic logic. I present an alternative way to handle prioritized norms in the setting of I/O logic, and show that while in simple settings both accounts coincide, this is not true in more complicated or "realistic" settings in which more than two norms collide. My aim is more general than just comparing two particular logical systems. I contrast two strategies of formalizing defeasible deontic reasoning. The first one, referred to as the "increment" idea, is well-known from the non-monotonic literature. It consists in restraining the step-by-step application of defaults when building the extension. A triggered rule is applied only if its consequent is consistent with what has previously been inferred. The second one, referred to as the "threshold" idea, is much less known. It consists in cutting back the set of defaults to just below the threshold of yielding excess. This is implemented using the notion of maxfamily mentioned above, and the notion of preffamily (this is short for "preferred family"). One starts with a reflexive and transitive priority relation $\geq$ on the members of N. One lifts it to a relation $\geq^s$ on subsets of N. One then sets:

- preffamily: $\mathit{preff}(N, A, C)$ is the set of all the Hs in $\mathit{outf}(N, A, C)$ that are maximal under $\geq^s$.

Intuitively, the preffamily lists the "preferred" elements in the maxfamily. The final output is restricted to these elements.

In [78], I give an example where the threshold approach gives an intuitively more satisfactory solution than the increment approach. The example was suggested to me by Marek Sergot. $(a, x) > (b, y)$ is a shorthand for $(a, x) \geq (b, y)$ and $(b, y) \not\geq (a, x)$.

Example 3.5.1 (Cancer [78])**.** *Assume we have*

$$\{(b, c), (a, b)), (a, \neg b)\} \text{ with } (b, c) > (a, b) > (a, \neg b)$$

a is for the set of data used to set up a treatment against cancer, b is for receiving chemo as per the protocol, and c is for keeping the patient's WBCs (White Blood Cells) count to a safe level. In a diagram:

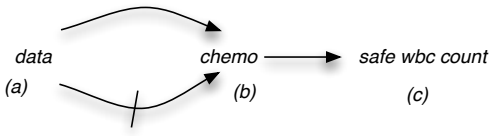

Figure 3.5.1: Cancer

The difference appears when the input is $\{a, \neg c\}$. In that case, the threshold account outputs the obligation of $\neg b$ while the increment account outputs the obligation of b. Following one of Hansson [49]'s suggestions, one should think of the input as settled as true. This means that c is out of reach. It is impossible for the doctor to maintain the patient's WBC count at a safe level using a drug, because s.he is too weak. The question is: is b obligatory or not? The ordering $(a, b) > (a, \neg b)$ says that (a, b) has priority over $(a, \neg b)$. So it would seem to follow that b is obligatory. The rationale behind the increment account, is that one just looks at the norms that are triggered, and one applies the stronger one. But, in reply, it can be said that the ordering $(b, c) > (a, \neg b)$ tells us that compliance with the stronger of the two conflicting norms triggers an obligation of even higher rank, namely the obligation to do c. Furthermore, c is already (settled as) false. Hence, if the agent makes b true, s.he will violate a norm with an even higher rank. The only way to prevent this violation of the most important norm is to make b false. Hence the answer returned by the threshold account. This aligns with what doctors typically do: if the WBC count cannot be maintained at a safe level, chemo is postponed until the patient has recovered.

Publications: [77, 78].

[77] received the best paper award at DEON 2010.

3.5.2 Proof systems and link with preferences.

The axiom systems for I/O logics incorporating a built-in consistency check serve as a "surrogate" proof theory for constrained I/O logic. In particular, one gets a purely proof-theoretic solution to the question of accommodating the existence of deontic dilemmas without generating deontic explosion. The earlier treatment with constrained I/O logic was purely semantical. A deontic dilemma is a situation where both an action a and its negation are obligatory. There is a deontic explosition, if

every action b is obligatory. Here is the derivation standardly used to show how deontic explosion arises:

$$\dfrac{\dfrac{(\top, a) \qquad (\top, \neg a)}{(\top, a \wedge \neg a)}\,\text{AND}}{(\top, b)}\,\text{WO}$$

One could think that the rule R-AND is enough to block the derivation of $(\top, b)$. However, this is not the case, as shown by van Fraassen [112]'s paradox:

$$\dfrac{\dfrac{\dfrac{(\top, a)}{(\top, a \vee b)}\,\text{WO} \qquad (\top, \neg a)}{(\top, (a \vee b) \wedge \neg a)}\,\text{R-AND}}{(\top, b)}\,\text{WO}$$

Van Fraassen's paradox.

This gives another independent reason for letting WO go.

The proof system is simpler (yet complementary) than the one given in [109], which relies on adaptive logic.

The connection to preference-based dyadic deontic logic is further discussed in Section 5.1.2.

3.6 Applications

I discuss two applications I have been working on: Kelsen's analysis of a norm; moral particularism.

3.6.1 Kelsen's analysis of a norm. Bochman [18] has argued that input/output logics represent the appropriate formalism for analyzing causality, because we do not have ID. In collaboration with A. Ciabattoni and G. Sartor, we envisaged a similar but different application. The idea is to reconstruct deontic logic in accordance with the analysis of norms proposed by the legal theorist Kelsen. According to him, a norm establishes a relationship of imputation between a transgression and a sanction, which is the legal analog of the notion of causal relationship in the empirical sciences. A norm imputes the sanction (punishment) to a given act (the act of theft). It does not describe a factual or causal link–it prescribes a legal consequence.

On the logic side, we proceed in two steps. First, (unconstrained) I/O logic is extended with multiple violation constants (for obligation) and multiple license constants (for permission). Second, the modalities of obligation and permission are defined using a reduction schema similar to the one proposed by Anderson. A state of affairs is a duty, if its negation leads (in the I/O logic) to a violation, and permitted if its realization leads (in the I/O logic) to a license.

Publications: [29, 30].

3.6.2 Moral particularism. Moral particularism is an influential theory in contemporary moral philosophy, put forth by Dancy [33]. My work takes for granted that the traditional view of ethical reasoning as principle-based (or top-down) is correct. Proponents of moral particularism have dismissed this view based on reason holism. According to reason holism, a consideration that is a reason in one context may not be similarly a reason in another context because of differences in the presence or absence of defeating and enabling conditions.

Moral particularism can be used in AI & Ethics to justify a purely bottom-up approach to AI ethics rooted in machine learning. Dancy [33, p. 111] himself suggests that connectionism may serve as a model for learning norms.

If particularists were right, my (our) research would be pointless. In response, I present two key arguments. First, I argue that contrary-to-duty obligations (CTDs) introduce a distinct source of reason holism. Second, elaborating on an idea proposed by Horty, I use the prioritized I/O logic defined above to block the inference from reason holism to moral particularism. The proposed framework supports reason holism: a reason need not retain its supporting value across contexts. Yet, it is based on a system of principles. These are thought of as defeasible generalizations– statements that are usualy true, but can have exceptions. The framework provides a counter-example to the claim that reason holism implies particularism.

Publications: [77, 78].

As mentioned, [77] received the best paper award at DEON 2010.

Chapter 4

Automated reasoning

I describe my contribution to automated normative reasoning. This is joint work with C. Benzmüller. Section 4.1 gives the motivation for this work and explains the method. Section 4.2 summarizes the results achieved. Section 4.3 gives the example of system **E**. Section 4.4 describes an application.

4.1 Motivation

There are two main approaches to automation: direct and indirect. In the direct approach, a native prover is developed specifically for a given logic, call it $\mathcal{L}_1$. In contrast, the indirect approach involves translating $\mathcal{L}_1$ (source logic) into another logic $\mathcal{L}_2$ (target logic), and then using an existing prover for $\mathcal{L}_2$ to automate reasoning tasks performed in $\mathcal{L}_1$.

Each approach has its own strengths and weaknesses. Automated theorem proving (ATP) systems have been developed for modal logic using either approach. By contrast, there has been limited work on developing ATP systems for conditional deontic logic. My work with C. Benzmüller aims at filling in this gap, applying the indirect method he has been developing with his team during the last 20 years, called Shallow Semantical Embedding (SSE). The target of the embedding is classical higher-order logic (HOL), *i.e.*, Church's type theory [6].

There are three main steps. The first step is to specify an embedding $\lfloor \cdot \rfloor$, which translates a formula A of the source logic into a formula $\lfloor A \rfloor$ of HOL. The second step is to establish that the embedding is sound and complete, that is faithful, in the

sense that it preserves both the validity and invalidity of formulas. The establishment of such a result is the main criterion of success. Such a result provides a guarantee the reasoner will use the source logic to answer queries. The third and last step consists in encoding the embedding in a concrete proof assistant. We worked with Isabelle/HOL [70]. It internally provides automated reasoning tools such as *Sledgehammer* and *Nitpick* [16, 17]. The automated theorem proving systems integrated via *Sledgehammer* include higher-order ATP systems, first-order ATP systems, and SMT (satisfiability modulo theories) solvers, and many of these systems in turn use efficient SAT solver technology internally.

Shallow semantic embedding is in contrast to so-called deep embedding that encodes the source logic formulae as uninterpreted data (usually an inductively defined datatype), and defines meta-theoretical notions such as interpretation and satisfiability as functions and predicates. The shallow embedding of modal logics into classical (typed) first-order logic is commonly referred to as the standard translation and widely known. The proposed method can be viewed as a variation on it.

On the HOL side, the following two primitive types are used: i for individuals (or possible worlds); o for the Boolean values. A variant of the standard semantics is used. It is called "generalized" or (after its inventor) "Henkin" semantics. This variant semantics leads to an axiomatizable version of higher-order logic, because the set of functions in a given model need not be complete. (See [51].)

The main ideas of the embedding into HOL are informally sketched in Section 4.3 using the example of system **E**.

4.2 Results

The following systems have been "implemented" by utilizing the SSE approach.

SDL: All logics from the modal logic cube, including logic KD, i.e. SDL [13]. These implementations scale for first-order and even higher-order extensions.

DDL: the DDL by Åqvist [3, 83] and the DDL by Carmo and Jones [23]. This one comes with a neighborhood semantics

Unconstrained I/O logic: The main challenge comes from the fact that the framework does not have a truth-functional semantics, but an operational one. The implementation in [15] (based on the master's thesis of P. Meder, which I co-supervised together with L. van der Torre and C. Benzmüller) is doubly indirect: it takes advantage of the fact that two unconstrained I/O operations

can themselves be embedded into modal logic. The implementation in [11] does not presuppose a modal translation, but is still preliminary, and has not been followed up.

In general, the supported queries are:

- Proving the validity of formula, or proving that a conclusion follows from a set of assumptions (via *Sledgehammer*)
- Disproving a formula or showing consistency by providing a model (via *Nitpick*).

An overview of our contribution was published in *Artificial Intelligence*. Individual embeddings were published in separate papers. A data set of the Isabelle encodings was published in *Data in brief* and *Archive of Formal Proofs*, the official repository maintained by the Isabelle developers. The theory files are available on-line at LogiKEy's repository (sub-directory "deonticlogic").

Critical assessment Overall, our evaluation indicates that the provers exhibit good responsiveness. For example, validating a formula in **E** typically requires no more than 7 milliseconds. We also observed that the selection of a back-end significantly influences performance, with zipper position generally outperforming alternatives.

Besides, each embedding comes with a proof of its faithfulness. This provides a guarantee that the reasoner effectively uses the logic. It also explains why in our experiments the model generated by *Nitpick* was always correct. The provers provide answers to queries, but also explanations. In particular, if a formula is not valid, a counter-model is outputted.

A comparative study with native provers, similar to the one in [108] must be left as a topic for future work. We are not aware of a similar automation of the systems studied in this project using other methods. A comparison with a prover for a related system (e.g. KLMLean 2.0, due to Giordano [42]) would already be beneficial. Based on the evaluation made for KLMLean 2.0 by their authors, one can predict the direct method will offer superior performance metrics. However, it also presents greater implementation challenges. A native prover is tied to a specific system. For instance, one could try to translate the rules of our sequent calculus for **E** in Prolog. But the calculus uses a $\mathcal{B}et$ modality, encoding the notion of maximality in the syntax, and would be tied to it. We do not have a sequent calculus for systems with a different notion of best, like optimality or strong maximality.

One notable advantage of our approach is its inherent flexibility. In the absence of a universally accepted deontic logic, it is crucial to give users the freedom to choose their preferred deontic logic, let them dynamically modify the logic and explore the resulting implications. This flexibility is illustrated in the case study presented in Section 4.4, where our proof assistant is used to test different hypotheses and modeling choices on the fly.

Publications: [12] (AIJ overview); [8, 88] (Data publication); [9, 10, 11, 87, 89] (individual embeddings).

[10] received the best paper award at DEON 2018.

4.3 Sample embedding

In this section the main ideas of the embedding into HOL are informally outlined, by taking the case of Åqvist's system **E** and its extensions.

The monadic fragment of second-order logic, which is known to be decidable, is the target logic. The notion of preferential model is simulated in HOL as follows: a possible world is identified with an individual (of type i). A propositional letter p is identified with a predicate constant P. Thus a formula is of type $i \to o$, where o is the classical type of truth values. Type $i \to o$ is here abbreviated as τ. The betterness relation is identified with a binary predicate constant r of type $i \to i \to o$.

The mapping $\lfloor \cdot \rfloor$ translates a formula A of **E** into a formula $\lfloor A \rfloor$ of HOL of type τ. The mapping is defined recursively in the usual way for Boolean and alethic formulas. The clause for $\bigcirc$ is distinctive. There is ample room for variation and significant flexibility. If one chooses to define best in terms of optimality, one puts:

$$\lfloor \bigcirc(B/A) \rfloor \;=\; \bigcirc_{\tau\to\tau\to\tau} \lfloor B \rfloor \lfloor A \rfloor$$

where $\bigcirc_{\tau\to\tau\to\tau}$ abbreviates the following formula of HOL:

$$\bigcirc_{\tau\to\tau\to\tau} \;= \lambda B_\tau.\lambda A_\tau.\lambda x_i.\forall w_i.((\lambda v_i.(A\,v \wedge (\forall y_i.(A\,y \to r_{i\to\tau} v\,y))))\,w \to B\,w)$$

Thus, the basic idea is to translate a formula directly according to its semantics. Here $\bigcirc(B/A)$ is encoded as (omitting types other than i)

$$\lambda x_i.((\forall w_i.(Aw \wedge (\forall y.(Ay \to rwy)) \to Bw))$$

Global validity (*vld*) of an embedded formula A of **E** in HOL is defined by the equation

$$\mathrm{vld}\,\lfloor A \rfloor = \forall z_i.\lfloor A \rfloor z$$

These definitions are hidden from the user, who can construct now deontic logic formulas involving $\bigcirc(B/A)$ and use them to prove theorems.

It can be shown that the embedding is faithful, in the sense given by Theorem 4.3.1. Intuitively, Theorem 4.3.1 says that a formula A in the language of **E** is valid in the class of all preference models (call this class $\mathcal{P}$) if and only if its translation $\lfloor A \rfloor$ in the language of HOL is valid in the class of so-called Henkin models (call it Hen).

Theorem 4.3.1: Faithfulness of the embedding

$$\models^{\mathcal{P}} A \text{ if and only if } \models^{\text{Hen}} \text{vld} \lfloor A \rfloor$$

Proof. The proof can be found in Benzmüller *et al.* [10]. The crux of the argument consists in relating preference models with Henkin models in a truth-preserving way. □

The theorem is shown to hold under a rule of interpretation in terms of optimality. It straightforwardly extends to a rule of interpretation in terms of maximality. We have not considered strong maximality yet.

[10] establishes a similar result for Carmo and Jones [23]'s DDL. It uses a neighborhood semantics, and different types of modalities. Therefore, the embedding is more complex.

For the extensions of **E**, we have studied the correspondence between axioms and semantic conditions as "extracted" by relevant soundness and completeness theorems reported in Chapter 2. Thus, "correspondence" is taken in the same (broad) sense that Hughes and Cresswell have in mind when they write:

> "D, T, K4, KB [are] produced by adding a single axiom to K and [...] in each case the system turns out to be characterised by [sound and complete wrt] the class of models in which [the accessibility relation] R satisfies a certain condition. When such a situation obtains–i.e. when a system K+α is characterised by the class of all models in which R satisfies a certain condition–we shall [...] say [...] that the wff α itself is characterised by that condition, or that the condition *corresponds* [their italics] to α." [56, p. 41]

This is different from correspondence theory in the sense of Sahlqvist [111]. Typically, Sahlqvist-style modal correspondence theory studies the equivalence between modal formulas and first-order formulas over Kripke frames via the so-called standard translation. The goal is to identify syntactic classes of modal formulas that can be shown to define first-order conditions on frames, and which are themselves computable via an algorithm. Correspondence theory in this sense has not been developed for preference-based dyadic deontic logic and conditional logic yet. This is in part due to the more complex form of the truth conditions for the conditional obligation operator.

The situation for conditional (deontic) logic is still slightly different from the one for traditional modal logic. In the latter setting, the full equivalence between the property of the accessibility relation and the modal axiom is verified by automated means [7]. In the former setting only the direction "property ⇒ axiom" is verified by automated means. To be more precise, what is verified is the fact that, if the property holds, then the axiom holds. What is not confirmed is the converse statement, that if the axiom holds then the property holds.

Figure 4.3.1 shows the encoding of the embedding of Åqvist's system **E** in Isabelle/HOL. Be aware that there is a slight change of notation compared to the previous section. For instance, formulas are represented by Greek letters. The type for the truth values is now written "bool". Type $i \to o$ is abbreviated σ. On line 4, a designated constant *aw* (of type i) for the actual world is introduced. This is needed to be able to verify the truth or falsity of a formula in a given world. On lines 5-11, the propositional connectives are introduced. On line 20, a designated predicate constant R for the betterness relation is introduced. On line 22, optimality is defined. On lines 26-28, the deontic operators are introduced.

By way of further illustration, Figure 4.3.2 shows the formalization of Chisholm's paradox. On line 4, the predicates *go* and *tell* are introduced. On lines 10-16, the relevant obligations and fact are formalized.

In Figure 4.3.3, line 27, the query "Are D1, D2, D3 and D4 consistent?" is run. To reduce the search space of the model finder Nitpick, some assumptions are made: we ask for a model with three worlds ("card=3"), and (on lines 24-25) the betterness relation is required to be reflexive, transitive, total and limited. *Nitpick* confirms consistency. The Henkin model is shown in Figure 4.3.4. The model is correct.

Publications: [10, 87, 88, 89].

```
theory E imports Main       (* Aqvist's System E: C. Benzmüller & X. Parent, 2019 *)
begin
typedecl i (*Possible worlds.*) type_synonym σ = "(i⇒bool)"
consts aw::i (*Actual world.*)
abbreviation etrue  :: "σ" ("⊤") where "⊤ ≡ λw. True"
abbreviation efalse :: "σ" ("⊥")  where "⊥ ≡ λw. False"
abbreviation enot :: "σ⇒σ" ("¬_"[52]53)  where "¬φ ≡ λw. ¬φ(w)"
abbreviation eand :: "σ⇒σ⇒σ" (infixr"∧"51) where "φ∧ψ ≡ λw. φ(w)∧ψ(w)"
abbreviation eor  :: "σ⇒σ⇒σ" (infixr"∨"50) where "φ∨ψ ≡ λw. φ(w)∨ψ(w)"
abbreviation eimp :: "σ⇒σ⇒σ" (infixr"→"49) where "φ→ψ ≡ λw. φ(w)⟶ψ(w)"
abbreviation eequ :: "σ⇒σ⇒σ" (infixr"↔"48) where "φ↔ψ ≡ λw. φ(w)⟷ψ(w)"

(*Possibilist--constant domain--quantification.*)
abbreviation eforall ("∀") where "∀Φ ≡ λw.∀x. (Φ x w)"
abbreviation eforallB (binder"∀"[8]9) where "∀x. φ(x) ≡ ∀φ"
abbreviation eexists ("∃") where "∃Φ ≡ λw.∃x. (Φ x w)"
abbreviation eexistsB (binder"∃"[8]9) where "∃x. φ(x) ≡ ∃φ"

abbreviation ebox :: "σ⇒σ" ("□") where "□ ≡ λφ w.  ∀v. φ(v)"
consts R :: "i⇒σ" (infixr "R" 70) (*Betterness relation, cf. def. of ○<_|_>.*)
abbreviation eopt  :: "σ⇒σ" ("opt<_>")
  where "opt<φ> ≡ (λv. ( (φ)(v) ∧ (∀x. ((φ)(x)  ⟶  v R x) )) )"
abbreviation esubset :: "σ⇒σ⇒bool" (infix "⊆" 53)
  where "φ ⊆ ψ ≡ ∀x. φ x ⟶ ψ x"
abbreviation econd  :: "σ⇒σ⇒σ" ("○<_|_>")
  where "○<ψ|φ> ≡  λw. opt<φ> ⊆ ψ"
abbreviation euncobl :: "σ⇒σ" ("○<_>")
  where "○<φ> ≡ ○<φ|⊤>"
```

Figure 4.3.1: System **E**

4.4 Application: Parfit's repugnant conclusion

The aim is to use the tool to facilitate a computer-assisted assessment of ethical arguments in philosophy, thereby refining our understanding of these arguments. A similar approach has been successfully applied in computational metaphysics to evaluate variations of Gödel's ontological proof [14].

This study deals with Parfit's repugnant conclusion, a well-known paradox in population ethics [99]. It states: For any possible population of at least ten billion people, all with a very high quality of life, there must be some much larger imaginable population whose existence, if other things are equal, would be better even though its members have lives that are barely worth living.

Our formalization focuses on the "mere addition paradox," a component of the Repugnant Conclusion. By encoding the relevant assumptions into Isabelle/HOL, we were able to examine the logical structure of the paradox and explore potential res-

```
theory Chisholm_Exav imports E   (*Christoph Benzmüller & Xavier Parent, 2019*)

begin (* Chisholm Example *)
consts go::σ tell::σ kill::σ

nitpick_params [user_axioms,expect=genuine,show_all,format=2]
                              (*settings for the model finder*)

(*It ought to be that Jones goes to assist his neighbors.*)
  abbreviation  "D1 ≡ ○<go>"
(*It ought to be that if Jones goes, then he tells them he is coming.*)
  abbreviation  "D2 ≡ ○<tell|go>"
(*If Jones doesn't go, then he ought not tell them he is coming.*)
  abbreviation  "D3 ≡ ○<¬tell|¬go>"
(*Jones doesn't go. (This is encoded as a locally valid statement.)*)
  abbreviation  "D4 ≡ ¬go"
```

Figure 4.3.2: Chisholm's set

```
abbreviation "olimitedness ≡ (∀φ. (∃x. (φ)x) ⟶ (∃x. opt<φ>x))"
abbreviation "reflexivity  ≡ (∀x. x R x)"
abbreviation "transitivity ≡ (∀x y z. (x R y ∧ y R z) ⟶ x R z)"
abbreviation "totality ≡ (∀x y. (x R y ∨ y  R x))"

 (* Consistency *)
lemma assumes "reflexivity" and "olimitedness"
         and "transitivity" and "totality"
    shows  "⌊(D1 ∧ D2 ∧ D3)⌋ ∧ ⌊D4⌋ι"
    nitpick [satisfy,card=3] (*Consistent? Yes*)
```

Figure 4.3.3: Query

olutions overlooked so far.

We have used the tool to assess a solution to the repugnant conclusion proposed by Tempkin [110] and others. It consists in dropping the assumption of the transitivity of "better than". Our main finding is that weakening the transitivity of the "better than" relation, rather than abandoning it entirely, might offer a less extreme approach to addressing the paradox. However, not all the candidate weakenings perform equally: quasi-transitivity and acyclicity are ok, but not the interval order condition.

This insight was made possible by the flexibility of the reasoner. The results of our experiments are presented in Table 4.4.1.The left-most column shows the constraint put on the betterness relation. The other columns show what happens when varying the truth conditions for the conditional obligation operator. $\exists\forall$ refers to the evaluation rule proposed by D. Lewis. It puts $\bigcirc(B/A)$ as true, when there is a $A \wedge B$-

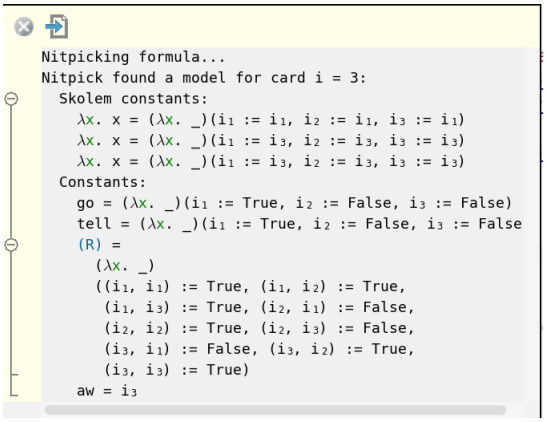

Figure 4.3.4: Henkin model for Chisholm's paradox.

world starting a (possibly infinite) sequence of increasingly better $A \to B$-worlds. The symbol ✓ indicates that the sentences formalizing the scenario have been confirmed to be consistent, and the symbol ✘ indicates they have been confirmed to be inconsistent.

Property \ Truth conditions	opt	max	∃∀
None	✓	✓	✓
Transitivity + totality	✘	✘	✘
Transitivity	✘	✘ (if model finite)	✘
Interval order	✘	✘	✘
Quasi-transitivity	✓	✘ (if model finite)	✓
Acyclicity	✓	✓	✓

Table 4.4.1: Mere addition paradox (experiments)

One can see that changing the truth conditions for the conditional does not have any effect, except for transitivity and quasi-transitivity. "Finite model" refers to the fact that the model has finitely many worlds. We discovered that under the max rule the finiteness of the model is needed to generate an inconsistency: without this assumption, the model finder times out. This observation had not been previously documented in the literature. It also led us to identify a case where the finite model property (f.m.p.) fails. This property asserts that if a formula is satisfiable, then it is satisfiable in a finite model.

Theorem 4.4.1: f.m.p., negative result

Under the max rule, the finite model property fails w.r.t. the following classes of models whose relation $\succeq$ meets the property as indicated:

- $\succeq$ is quasi-transitive
- $\succeq$ is transitive
- $\succeq$ is an interval order

Proof. By showing (manually) that the conjunction of the formulas formalizing the mere addition paradox is not satisfiable in a finite model in the considered class. □

Previously, all these points could have been verified manually using pen and paper. Now, computers can perform this task. Furthermore, while the proof assistant did not autonomously establish Theorem 4.4.1, our interaction with it triggered its discovery.

Publication: [88, 89].

Chapter 5

Projects

5.1 Short-term: continuation of current research

5.1.1 Preference-based dyadic deontic logic

I am interested in the following three topics:

- To achieve analogous completeness results for variant evaluation rules for the conditional obligation operator, proposed to address deficiencies when no best world exists
- To extend the account to defeasible conditional obligation, and make a formal bridge with norm-based dyadic deontic logic (input/output logic)
- To investigate issues raised by the extension to the first-order case

The next sections discuss these points.

Variant truth-conditions

One variant I am interested in is the $\exists\forall$ rule proposed by D. Lewis. It puts $\bigcirc(B/A)$ as true, when there is a $A \wedge B$-world starting a (possibly infinite) sequence of increasingly better $A \rightarrow B$-worlds. Lewis [61] (among others) axiomatized the class of models where the betterness relation comes with the full panoply of all the standard properties (except the limit assumption in any of its forms). His system VTA is very close to Åqvist's system **G**. I am interested in relaxing these properties.

To let totality go allows us to accommodate the possibility of unresolved conflicts between obligations, viz situations where $\bigcirc(B/A)$ and $\bigcirc(\neg B/A)$ both hold. The

axiomatization problem for the partial order case was resolved by Goble [44]. I am interested in understanding if his result carries over to the setting I have been working with, and what minimal changes must be made to **G** to obtain an analog result. The assumptions of reflexivity and transitivity of the betterness relation are both still needed. I would like to understand if it is possible to relaxe these two properties, in particular transitivity. I also would like to understand if the weaker forms of transitivity discussed in Section 2.2 have a syntactical counterpart in this setting.

Defeasible reasoning

This (on-going) work aims at addressing a concern raised by Horty (in e.g. [54, 55]) regarding the ability of the preference-based approach (as described in Chapter 2) to model defeasible reasoning. A reasoning is defeasible, when conclusions can be retracted in the light of further information. This type of reasoning typically involves obligations than can be overridden by more specific obligations.

I discuss Horty's objections and then present a semantic account designed to address them. His main objection is that the preference-based approach to conditional obligation was developed as a species of modal logic, within the framework of possible worlds semantics. The consequence relation is monotonic: when the premise set grows, the set of conclusions also grows. This makes them unsuitable to model defeasible reasoning.

Nonmonotonicity of the consequence relation is obtained, by parametrizing this one by a set of conditionals, referring to e.g. the normative system under consideration. The idea is not new, and was studied in nonmonotonic logic before. The first innovation is the use of a bi-ordering semantics, with two types of rankings in the models: one based on ideality ($\succeq_I$) and the other based on normality ($\succeq_N$). This allows to resolve problems pointed out raised by Horty and other deontic logicians, such as the fallacy of the prohibited exception. The fallacy of the prohibited exception is the derivation of $\bigcirc\neg a$ ("You ought not to eat asparagus!") from $\bigcirc\neg f$ ("You ought not to eat with your fingers") and $\bigcirc(f/a)$ ("You ought to eat with your fingers when you eat asparagus"). This derivation holds in Åqvist's systems and most existing preference-based dyadic deontic logics. It makes these frameworks unsuitable for reasoning about exceptions.

The norms in the normative system, along with the conditionals in the premise set, are used to further constrain the ranking within the model. Different methods are applied to each type of conditional: for ideality, I use Delgrande's method [35],

while for normality, I use Lehmann's method [59]. In a second step, to give it more strength, the account is extended to support conflict resolution through a partial ordering $>$ on the set of obligations. Intuitively, $\bigcirc(B/A) > \bigcirc(D/C)$ says that the first obligation takes precedence over the second. Here the main technical challenge is to understand how the ordering $>$ on the obligations affects the ordering on the possible worlds in the model.

The expected added value is a better understanding of the interplay between obligation statements and normality statements, statements about what is typically the case.

Publication: [32] and [86] (under revision).

First-order deontic reasoning

I summarize preliminary results obtained in the first-order case. This is joint work with D. Pichler, whose master I co-supervised, and whose PhD I co-supervise (with A. Ciabattoni).

Little work has been done in general on first-order deontic logic. One reason for this is the widespread belief that the extension of propositional deontic logic to first-order systems does not raise any new interesting issues, and follows the pattern already established in other areas of modal logic. Pace this common belief, Goble has argued that new issues arise in the deontic case, particularly concerning the notion of extensionality.

An operator is extensional if it allows salva veritate substitution of co-referential terms, and intensional if it does not. The distinctiveness of the "ought" modality lies in its extensionality. That is, we content that the principle of substitution salva veritate of co-referential terms should hold for the "ought" modality. On the other hand, it is known that its addition leads to the deontic collapse, viz $A \equiv \bigcirc A$ becomes a theorem.

We explore whether extensional and intensional operators can coexist within the same semantics without causing this deontic collapse, using Åqvist's system **F** for conditional obligation. This is the weakest system in which the collapse arises. The original semantics for **F** is extended to first-order logic with definite descriptions, and made two-dimensional. The basic idea is to assume that terms withing the scope of "ought" take the reference they have in the actual world. The framework offers a more nuanced approach to first-order deontic principles by employing a cross-world, perspectival evaluation.

A proper name is usually considered a rigid designator: its reference does not vary from one world to another. It is for definite descriptions that the problem of extensionality arises. A definite description is a phrase of the form "the ϕ". Its reference is generally considered non-rigid, varying from one world to another. We build on the insights of Donnellan, Kaplan, and others, who, while accepting Kripke's main argument about proper names being rigid designators, observed that certain uses of definite descriptions appear to be *directly referential*: its meaning lies in what it points out in the world. Thus, the question is: how to account for the validity of the principle of substitution salva veritate, when a term is a definite description used this way? And this, without creating the deontic collapse.

Publications: [91, 100]

[100] received the best paper award at DEON 2023.

5.1.2 Towards a synthesis with norm-based deontic logic

I'm currently working on a synthesis between preference-based dyadic deontic logics and norm-based deontic logic (input/output logics). This is on-going work.

The deontic and flat fragment of the logic described in Section 5.1.1 ("Defeasible reasoning") has been faithfully embedded into so-called constained input/output logic. I briefly describe the embedding below. ("Flat" means that all obligations are considered equally, viz., the priority relation on the obligations is empty.)

On the preference side, the Hansson-Lewis semantics is extended by a set of conditions on the preference orderings intended to ensure that worlds are ranked according to how well they comply with explicitly given obligations, ignoring non-compliance to obligations that are overriden by more specific ones. A preference model is thus parametrized by a set of explicit obligations, call it N.

On the I/O side, the basic idea is to rewrite $\bigcirc(x/a)$ as an I/O pair with a more fine-grained antecedent, containing the negation of the antecedents of the obligations overriding it. For instance, if $\bigcirc(x/a)$ and $\bigcirc(\neg x/a \wedge b)$ are in N, then the second obligation overrides the first one. This one is then rewritten as $(a \wedge \neg(a \wedge b), x)$, or equivalently $(a \wedge \neg b, x)$. Intuitively, this amounts to requiring that x is obligatory if and only in so far as the condition b whose presence would trigger the overriding obligation is false.

Let N be a set of dyadic obligations, and $N^{\triangleright}$ be the result of rewriting the obligations in N as just described. The main result is that $\bigcirc(x/a)$ holds in the preference model

whose ideality relation $\succeq$ on worlds is determined by N if and only if x is in the so-called full meet constrained output under $N^{\triangleright}$ given input a, with the reusable throughout operation out_4^+ as the underlying I/O operation. Formally,

Theorem 5.1.1: Faithfullness of the embedding

$w \models \bigcirc(x/a)$ iff $x \in \cap outf(N^{\triangleright}, a, a)$, with out_4^+ as the background I/O operation.

The following extensions are topics for future research:

- Prioritized I/O logic
- Permission
- Alternative norm-based systems, like Horty [54]'s default logic.

Publication: [86] (under revision).

5.2 Long-term: utilitarian conditional deontic logic

This project contributes to the field of normative reasoning and ethical theory by developing a general framework for utilitarian conditional deontic logics—logical systems that model obligations from a utilitarian perspective. It builds on preference-based dyadic deontic logics, particularly those described in Chapter 2, and pursues further the planned integration with norm-based systems described in the previous section.

The project investigates how preference-based logics relate to ethical theories, focusing on one of the most prominent one, utilitarianism. An action is obligatory if it leads to the best possible outcome.

The project is guided by two key hypotheses:

1. Utilitarian goodness is quantitative–it can be measured rather than just ranked qualitatively.

2. Utilitarian reasoning is context-sensitive and defeasible–conclusions can be revised when new information comes, allowing for exceptions or conflicts between obligations.

The research addresses three main questions:

1. Numerical semantics: Which dyadic deontic systems can incorporate utility functions, even when the assumptions traditionally associated with a utility function are dropped, like the transitivity of "better than"?
2. Axiomatization and automation: Existing systems are for a simple form of utilitarianism, so-called act-utilitarianism (which focuses on individual actions). How can more sophisticated versions of utilitarianism be formalized, axiomatized and automatized? This includes utilitarianism based on expected utility, rule-utilitarianism (which focuses on the consequences of following a rule) and justice-adjusted utilitarianism (which aligns utilitarianism with common-sense morality)?
3. Defeasibility: How can these systems be extended to support defeasible reasoning?

The project's broader goal is to understand which conditional deontic logics are best suited for utilitarian reasoning, and which apply better to other moral frameworks.

Methodologically, it combines semantic and axiomatic approaches (from modal logic), applies measurement theory, and includes case studies. Addressing defeasibility involves unifying the two main traditions in deontic logic: preference-based and norm-based systems.

Although the focus is on a specific ethical theory, I am not aware of any comprehensive investigation into the interplay between logical structures and ethical theories. I consider such an inquiry to be of significant relevance to the field of AI ethics.

Bibliography

[1] Recueil de Législation Routière. http://legilux.public.lu/eli/etat/leg/code/route/20190531. Accessed: 2019-06-12.

[2] L. Åqvist. *An Introduction to Deontic logic and the Theory of Normative Systems*. Bibliopolis, Naples, 1987.

[3] L. Åqvist. Deontic logic. In D. Gabbay and F. Guenthner, editors, *Handbook of Philosophical Logic*, volume 8, pages 147–264. Kluwer Academic Publishers, Dordrecht, Holland, 2nd edition, 2002. Originally published in [38, pp. 605–714].

[4] O. Arieli, K. van Berkel, and C. Straßer. Defeasible normative reasoning: A proof-theoretic integration of logical argumentation. In M. J. Wooldridge, J. G. Dy, and S. Natarajan, editors, *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2014, February 20-27, 2024, Vancouver, Canada*, pages 10450–10458. AAAI Press, 2024.

[5] Z. Baniasadi, X. Parent, C. Max, and M. Cramer. A model for regulating of ethical preferences in machine ethics. In *Proceedings of HCI 2018*, pages 481–506. Springer, 2018. Best paper award.

[6] C. Benzmüller and P. Andrews. Church's type theory. In E. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, summer 2019 edition, 2019. https://plato.stanford.edu/entries/type-theory-church/.

[7] C. Benzmüller, M. Claus, and N. Sultana. Systematic verification of the modal logic cube in Isabelle/HOL. In C. Kaliszyk and A. Paskevich, editors, *PxTP 2015*, volume 186, pages 27–41, Berlin, Germany, 2015. EPTCS.

[8] C. Benzmüller, A. Farjami, D. Fuenmayor, P. Meder, X. Parent, A. Steen, L. van der Torre, and V. Zahoransky. Logikey workbench: Deontic logics, logic combinations and expressive ethical and legal reasoning (Isabelle/HOL dataset). *Data in Brief*, 33(106409):1–10, 2020.

[9] C. Benzmüller, A. Farjami, and X. Parent. A dyadic deontic logic in HOL. In J. Broersen, C. Condoravdi, S. Nair, and G. Pigozzi, editors, *Deontic Logic and Normative Systems – 14th International Conference, DEON 2018*, pages 33–50. College Publications, 2018. John-Jules Meyer Best Paper Award.

[10] C. Benzmüller, A. Farjami, and X. Parent. Åqvist's dyadic deontic logic E in HOL. *Journal of Applied Logics*, 6(5):733–755, 2019.

[11] C. Benzmüller and X. Parent. I/O logic in HOL – first steps. Technical report, CoRR, 2018. https://arxiv.org/abs/1803.09681.

[12] C. Benzmüller, X. Parent, and L. van der Torre. Designing normative theories for ethical and legal reasoning: Logikey framework, methodology, and tool support. *Artificial Intelligence*, 287:103348, 2020.

[13] C. Benzmüller and L. Paulson. Quantified multimodal logics in simple type theory. *Logica Universalis (Special Issue on Multimodal Logics)*, 7(1):7–20, 2013.

[14] C. Benzmüller and B. Woltzenlogel Paleo. The inconsistency in Gödel's ontological argument: A success story for AI in metaphysics. In S. Kambhampati, editor, *IJCAI 2016*, volume 1-3, pages 936–942. AAAI Press, 2016.

[15] C. Benzmüller, A. Farjami, P. Meder, and X. Parent. I/O logic in HOL. *Journal of Applied Logics*, 6:715–732, 2019.

[16] J. C. Blanchette, S. Böhme, and L. C. Paulson. Extending Sledgehammer with SMT solvers. *Journal of Automated Reasoning*, 51(1):109–128, 2013.

[17] J. C. Blanchette and T. Nipkow. Nitpick: A counterexample generator for higher-order logic based on a relational model finder. In Matt Kaufmann and Lawrence C. Paulson, editors, *ITP 2010*, volume 6172 of *LNCS*, pages 131–146. Springer, 2010.

[18] A. Bochman. *Explanatory Nonmonotonic Reasoning*. World Scientific, 2005.

[19] P. Boghossian. Knowledge of logic. In P. Boghossian and C. Peacocke, editors, *New Essays on the A Priori*, pages 229–254. Clarendon Press, Oxford, 2000.

[20] R. Booth and I. Varzinczak. Conditional inference under disjunctive rationality. In K. Leyton-Brown and Mausam, editors, *Proceedings of the 35th AAAI Conference on Artificial Intelligence*. AAAI, 2021. Proofs in an unpublished manuscript.

[21] R. Bradley. A note on incompleteness, transitivity and Suzumura consistency. In C. Binder, G. Codognato, M. Teschl, and Y. Xu, editors, *Individual and Collective Choice and Social Welfare: Essays in Honour of Nick Baigent*. Springer, 2015.

[22] J. Broome. *Rationality Through Reasoning*. Wiley-Blackwell, 2013.

[23] J. Carmo and A. J. I. Jones. Completeness and decidability results for a logic of contrary-to-duty conditionals. *Journal of Logic and Computation*, 23(3):585–626, 2013.

[24] R. Chang. The possibility of parity. *Ethics*, 112:659–688, 2002.

[25] B. Chellas. *Modal Logic*. Cambridge University Press, Cambridge, 1980.

[26] R. Chisholm. Contrary-to-duty imperatives and deontic logic. *Analysis*, 24:33–36, 1963.

[27] A. Ciabattoni, N. Olivetti, and X. Parent. Dyadic obligations: Proofs and countermodels via hypersequents. In R. Aydogan et al., editors, *PRIMA 2022: Principles and Practice of Multi-Agent Systems - 24th International Conference*, volume 13753 of *Lecture Notes in Computer Science*, pages 54–71. Springer, 2022.

[28] A. Ciabattoni, N. Olivetti, X. Parent, R. Ramanayake, and D. Rozplokhas. Analytic proof theory for åqvist's system F. In J. Maranhão, C. Peterson, C. Straßer, and L. van der Torre, editors, *Deontic Logic and Normative Systems - 16th International Conference, DEON 2023*, pages 79–98. College Publications, 2023.

[29] A. Ciabattoni, X. Parent, and G. Sartor. A Kelsenian deontic logic. In E. Schweighofer, editor, *Legal Knowledge and Information Systems - JURIX 202*, volume 346, pages 141–150. IOS Press, 2021.

[30] A. Ciabattoni, X. Parent, and G. Sartor. Permission in a Kelsenian perspective. In G. Sileno, J. Spanakis, and G. van Dijck, editors, *Legal Knowledge and Information Systems - JURIX 2023*, volume 379, pages 113–118. IOS Press, 2023.

[31] A. Ciabattoni and D. Rozplokhas. Streamlining input/output logics with sequent calculi. In P. Marquis, T. Cao Son, and G. Kern-Isberner, editors, *Proceedings of the 20th International Conference on Principles of Knowledge Representation and Reasoning, KR 2023*, pages 146–155, 2023.

[32] A. Ciabattoni, B. Istenic Urh, and X. Parent. Explanations in Åqvist's systems, 2025. Forthcoming in DEON 2025 proceedings.

[33] J. Dancy. *Ethics Without Principles*. OUP, 2004.

[34] S. Danielsson. *Preference and Obligation, Studies in the Logic of Ethics*. Filosofiska Fareningen, 1968.

[35] J. P. Delgrande. A preference-based approach to defeasible deontic inference. In D. Calvanese, E. Erdem, and M. Thielscher, editors, *Proceedings of KR 2020*, pages 326–335, 2020.

[36] M. Freund. A semantic characterization of disjunctive relations. In P. Jorrand and J. Kelemen, editors, *Fundamentals of Artificial Intelligence Research*, pages 71–83, Berlin, Heidelberg, 1991. Springer Berlin Heidelberg.

[37] N. Friedman and J. Y. Halpern. On the complexity of conditional logics. In J. Doyle, E. Sandewall, and P. Torasso, editors, *Proceedings of the 4th International Conference on Principles of Knowledge Representation and Reasoning (KR'94)*, pages 202–213, Cambridge, Massachusetts, 1994. Morgan Kaufmann.

[38] D. Gabbay and F. Guenthner, editors. *Handbook of Philosophical Logic*, volume II. Reidel, Dordrecht, Holland, 1st edition, 1984.

[39] D. Gabbay, J. Horty, X. Parent, R. van der Meyden, and L. van der Torre, editors. *Handbook of Deontic Logic and Normative Systems*, volume 1. College Publications, London, 2013.

[40] D. Gabbay, J. Horty, X. Parent, R. van der Meyden, and L. van der Torre, editors. *Handbook of Deontic Logic and Normative Systems*, volume 2. College Publications, London. UK, 2021.

[41] L. Giordano, V. Gliozzi, N. Olivetti, and G. L. Pozzato. Analytic tableaux calculi for KLM logics of nonmonotonic reasoning. *ACM Transactions on Compututational Logic*, 10(3):1–47, 2009.

[42] L. Giordano, V. Gliozzi, and G. L. Pozzato. KLMLean 2.0: A theorem prover for KLM logics of nonmonotonic reasoning. In N. Olivetti, editor, *Automated*

Reasoning with Analytic Tableaux and Related Methods, pages 238–244, Berlin, Heidelberg, 2007. Springer.

[43] L. Goble. A logic of good, should, and would: Part I. *Journal of Philosophical Logic*, 19:169–199, 1990.

[44] L. Goble. Preference semantics for deontic logics. Part I: Simple models. *Logique & Analyse*, 46(183-184):383–418, 2003.

[45] D. Grossi, W. van der Hoek, and L. B. Kuijer. Reasoning about general preference relations. *Artificial Intelligence*, 313(C), 2022.

[46] C. Hamblin. *Imperatives*. Blackwell, Oxford, 1987.

[47] J. Hansen. Prioritized conditional imperatives: problems and a new proposal. *Autonomous Agents Multi Agent Systems*, 17(1):11–35, 2008.

[48] J. Hansen. Reasoning about permission and obligation. In Hansson [50], pages 287–333.

[49] B. Hansson. An analysis of some deontic logics. *Noûs*, 3(4):373–398, 1969. Reprinted in [52, pp. 121-147].

[50] S. O. Hansson, editor. *David Makinson on Classical Methods for Non-Classical Problems*. Springer Netherlands, Dordrecht, 2014.

[51] L. Henkin. Completeness in the theory of types. *Journal of Symbolic Logic*, 15(2):81–91, 06 1950.

[52] R. Hilpinen, editor. *Deontic Logic: Introductory and Systematic Readings*. Reidel, Dordrecht, 1971.

[53] J. Horty. Defaults with priorities. *Journal of Philosophical Logic*, 36(4):367–413, 2007.

[54] J. Horty. *Reasons as Defaults*. Oxford University Press, 2012.

[55] J. Horty. Deontic modals: Why abandon the classical semantics? *Pacific Philosophical Quarterly*, 95(4):424–460, 2014.

[56] G. E. Hughes and M. J. Cresswell. *A Companion to Modal Logic*. Methuen, London, 1984.

[57] S. Kraus, D. Lehmann, and M. Magidor. Nonmonotonic reasoning, preferential models and cumulative logics. *Artificial Intelligence*, 44(1-2):167–207, 1990.

[58] D. Lassiter. *Graded Modality*. Oxford University Press, 2017.

[59] D. Lehmann. Another perspective on default reasoning. *Ann. Math. Artif. Intell.*, 15(1):61–82, 1995.

[60] D. Lehmann and M. Magidor. What does a conditional knowledge base entail? *Artificial Intelligence*, 55(1):1–60, 1992.

[61] D. Lewis. *Counterfactuals*. Blackwell, Oxford, 1973.

[62] P. Livet and X. Parent. Argumentation, révision et conditionnel. In P. Livet, editor, *Révision des Croyances*, pages 229–258. Hermès, Paris, 2001.

[63] R. Luce. Semiorders and a theory of utility discrimination. *Econometrica*, 24:178–191, 1956.

[64] D. Makinson. General theory of cumulative inference. In M. Reinfrank, J. de Kleer, M. Ginsberg, and E. Sandewall, editors, *Non-Monotonic Reasoning, 2nd International Workshop, Grassau, FRG, June 13-15, 1988, Proceedings*, volume 346 of *Lecture Notes in Computer Science*, pages 1–18. Springer, 1988.

[65] D. Makinson. Five faces of minimality. *Studia Logica*, 52(3):339–379, 1993.

[66] D. Makinson. General patterns in nonmonotonic reasoning. In D. Gabbay, C.J. Hogger, and J.A. Robinson, editors, *Handbook of Logic in Artificial Intelligence and Logic Programming*, volume 3. Oxford University Press, 1994.

[67] D. Makinson and L. van der Torre. Input/output logics. *Journal of Philosophical Logic*, 29(4):383–408, 2000.

[68] D. Makinson and L. van der Torre. Constraints for input/output logics. *Journal of Philosophical Logic*, 30(2):155–185, 2001.

[69] D. Makinson and L. van der Torre. Permission from an input/output perspective. *Journal of Philosophical Logic*, 32(4):391–416, Aug 2003.

[70] T. Nipkow, L.C. Paulson, and M. Wenzel. *Isabelle/HOL: A Proof Assistant for Higher-Order Logic*, volume 2283 of *LNCS*. Springer, Lecture Notes in Computer Science, 2002.

[71] M. Olszewski, X. Parent, and L. van der Torre. Input/output logic with a consistency check - the case of permission. In F. Liu, A. Marra, P. Portner, and F. Van De Putte, editors, *Deontic Logic and Normative Systems - 15th In-*

ternational Conference, DEON 2020/21, pages 358–375. College Publications, 2021.

[72] M. Olszewski, X. Parent, and L. Van der Torre. Permissive and regulative norms in deontic logic. *Journal of Logic and Computation*, 34(4):728–763, 2023.

[73] X. Parent. Cumulativity, identity and time in deontic logic. *Fundamenta Informaticae*, 48(2-3):237–252, 2001.

[74] X. Parent. *Nonmonotonic Logics and Modes of Argumentation*. PhD thesis, University of Aix-Marseille, 2002.

[75] X. Parent. Remedial interchange, contrary-to-duty obligation and commutation. *J. Appl. Non Class. Logics*, 13(3-4):345–375, 2003.

[76] X. Parent. On the strong completeness of Åqvist's dyadic deontic logic G. In R. van der Meyden and L. van der Torre, editors, *Deontic Logic in Computer Science, 9th International Conference, DEON 2008*, volume 5076 of *Lecture Notes in Computer Science*, pages 189–202. Springer, 2008.

[77] X. Parent. Moral particularism and deontic logic. In G. Governatori and G. Sartor, editors, *Deontic Logic in Computer Science, 10th International Conference, DEON 2010*, volume 6181 of *Lecture Notes in Computer Science*, pages 84–97. Springer, 2010.

[78] X. Parent. Moral particularism in the light of deontic logic. *Artificial Intelligence and Law*, 19(2-3):75–98, 2011.

[79] X. Parent. Why be afraid of identity? In A. Artikis, R. Craven, N. K. Cicekli, B. Sadighi, and K. Stathis, editors, *Logic Programs, Norms and Action - Essays in Honor of Marek J. Sergot on the Occasion of His 60th Birthday*, volume 7360 of *Lecture Notes in Artificial Intelligence*, Heidelberg, 2012. Springer.

[80] X. Parent. Maximality vs. optimality in dyadic deontic logic. *Journal of Philosophical Logic*, 43(6):1101–1128, 2014.

[81] X. Parent. Completeness of Åqvist's systems E and F. *Review of Symbolic Logic*, 8(1):164–177, 2015.

[82] X. Parent. A modal characterisation of an intuitionistic I/O operation. *Journal of Applied Logics*, 8(8):2349–2362, 2021.

[83] X. Parent. Preference semantics for dyadic deontic logic: a survey of results. In Gabbay et al. [40], pages 7–70.

[84] X. Parent. On some weakened forms of transitivity in the logic of norms. In O. Arieli, G. Casini, and L. Giordano, editors, *Proceedings of the 20th International Workshop on Non-Monotonic Reasoning, NMR@Floc*, volume 3197 of *CEUR Workshop Proceedings*, pages 147–150. CEUR-WS.org, 2022.

[85] X. Parent. On some weakened forms of transitivity in the logic of conditional obligation. *Journal of Philosophical Logic*, 53(3):721–760, 2024.

[86] X. Parent. On a problem of J. Horty. Under review, 2025.

[87] X. Parent and C. Benzmüller. Automated verification of deontic correspondences in Isabelle/HOL - first results. In C. Benzmüller and J. Otten, editors, *Proceedings of the 4th International Workshop on Automated Reasoning in Quantified Non-Classical Logics (ARQNL 2022)*, volume 3326 of *CEUR Workshop Proceedings*, pages 92–108. CEUR-WS.org, 2022.

[88] X. Parent and C. Benzmüller. Conditional normative reasoning as a fragment of HOL (Isabelle/HOL dataset). *Archive of Formal Proofs*, March 2024. https://isa-afp.org/entries/CondNormReasHOL.html, Formal proof development.

[89] X. Parent and C. Benzmüller. Normative conditional reasoning as a fragment of HOL. *Journal of Applied Non-Classical Logics*, 34(4):561–592, 2024.

[90] X. Parent, D. Gabbay, and L. van der Torre. Intuitionistic basis for input/output logic. In Hansson [50], pages 263–286.

[91] X. Parent and D. Pichler. Extensionality vs. intensionality in first-order deontic logic: A perspectival account. *Journal of Applied Logics*, 12:125–158, 2025.

[92] X. Parent and L. van der Torre. Input/output logic. In Gabbay et al. [39], pages 499–544.

[93] X. Parent and L. van der Torre. Aggregative deontic detachment for normative reasoning. In C. Baral, G. De Giacomo, and T. Eiter, editors, *Principles of Knowledge Representation and Reasoning: Proceedings of the Fourteenth International Conference, KR 2014*. AAAI Press, 2014.

[94] X. Parent and L. van der Torre. "Sing and dance!" - Input/output logics without weakening. In F. Cariani et al., editors, *Deontic Logic and Normative Systems - 12th International Conference, DEON 2014*, volume 8554 of *Lecture Notes in Computer Science*, pages 149–165. Springer, 2014.

[95] X. Parent and L. van der Torre. The pragmatic oddity in norm-based deontic logics. In G. Governatori, editor, *Proceedings of the 16th Edition of the International Conference on Articial Intelligence and Law*, ICAIL '17, pages 169–178, New York, NY, USA, 2017. Association for Computing Machinery.

[96] X. Parent and L. van der Torre. *Introduction to Deontic Logic and Normative Systems*. College Publications, 2018.

[97] X. Parent and L. van der Torre. I/O logics with a consistency check. In J. Broersen, C. Condoravdi, N. Shyam, and G. Pigozzi, editors, *Deontic Logic and Normative Systems - 14th International Conference, DEON 2018*, pages 285–299. College Publications, 2018.

[98] X. Parent and L. van der Torre. I/O logics without weakening. *Filosofiska Notiser*, 6:189—-208, 2019.

[99] D. Parfit. *Reasons and Persons*. OUP, 1984.

[100] D. Pichler and X. Parent. Perspectival obligation and extensionality in an alethic-deontic setting. In J. Maranhão, C. Peterson, C. Straßer, and L. van der Torre, editors, *Deontic Logic and Normative Systems - 16th International Conference, DEON 2023*, pages 57–77. College Publications, 2023. John-Jules Meyer best paper award.

[101] H. Prakken and M. Sergot. Dyadic deontic logic and contrary-to-duty obligations. In D. Nute, editor, *Defeasible Deontic Logic*, pages 223–262. Kluwer Academic Publishers, Dordrecht, 1997.

[102] H. Rott. *Change, Choice and Inference*. Clarendon Press, Oxford, 2001.

[103] P. A. Samuelson. A note on the pure theory of consumer's behavior. *Economica*, 5(17):61–71, 1938.

[104] K. Schlechta. *Nonmonotonic Logics, Basic Concepts, Results, and Techniques*, volume 1187 of *Lecture Notes in Computer Science*. Springer, 1997.

[105] Y. Shoham. *Reasoning About Change: Time and Causation from the Standpoint of Artificial Intelligence*. MIT Press, Cambridge, 1988.

[106] H. Simon. *Models of Man, Social and Rational*. John Wiley and Sons, New York, 1957.

[107] W. Spohn. An analysis of Hansson's dyadic deontic logic. *Journal of Philosophical Logic*, 4(2):237–252, 1975.

[108] A. Steen, G. Sutcliffe, T. Scholl, and C. Benzmüller. Solving modal logic problems by translation to higher-order logic. In A. Herzig, J. Luo, and P. Pardo, editors, *Logic and Argumentation*, pages 25–43, Cham, 2023. Springer Nature Switzerland.

[109] C. Strasser, M. Beirlaen, and F. Van De Putte. Adaptive logic characterizations of input/output logic. *Studia Logica*, 104(5):869–916, 2016.

[110] L. S. Temkin. Intransitivity and the mere addition paradox. *Philosophy and Public Affairs*, 16(2):138–187, 1987.

[111] J. van Benthem. Correspondence theory. In D. M. Gabbay and F. Guenthner, editors, *Handbook of Philosophical Logic*, pages 325–408. Springer Netherlands, Dordrecht, 2001.

[112] L. van der Torre and X. Parent. Detachment in normative systems: Examples, inference patterns, properties. *FLAP*, 9(4):1087–1130, 2022.

[113] B. C. van Fraassen. The logic of conditional obligation. *Journal of Philosophical Logic*, 1(3/4):417–438, 1972.

[114] B. C. van Fraassen. Values and the heart's command. *The Journal of Philosophy*, 70(1):5–19, 1973.