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 norma-

tive 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 Rosati [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