Informatics

A Logical Analysis of Normative Reasoning: Agency, Action, and Argumentation

DISSERTATION

submitted in partial fulfillment of the requirements for the degree of

Doktor der Technischen Wissenschaften

by

Cornelis Lambertus Johannes (Kees) van Berkel, BA MSc
Registration Number 11743091

to the Faculty of Informatics

at the TU Wien

Advisor: Prof. Dr. Agata Ciabattoni
Second advisor: Prof. Dr. Stefan Woltran

The dissertation has been reviewed by:

Ofer Arieli

Jan Broersen

[redacted]

Vienna, 27th February, 2023

Cornelis Lambertus Johannes
(Kees) van Berkel

Technische Universität Wien
A-1040 Wien ▪ Karlsplatz 13 ▪ Tel. +43-1-58801-0 ▪ www.tuwien.at

Erklärung zur Verfassung der Arbeit

Cornelis Lambertus Johannes (Kees) van Berkel, BA MSc

Hiermit erkläre ich, dass ich diese Arbeit selbständig verfasst habe, dass ich die verwendeten Quellen und Hilfsmittel vollständig angegeben habe und dass ich die Stellen der Arbeit – einschließlich Tabellen, Karten und Abbildungen –, die anderen Werken oder dem Internet im Wortlaut oder dem Sinn nach entnommen sind, auf jeden Fall unter Angabe der Quelle als Entlehnung kenntlich gemacht habe.

[redacted]

Wien, 27. Februar 2023

Cornelis Lambertus Johannes
(Kees) van Berkel

Acknowledgements

This thesis could not have been without my advisor Agata Ciabattoni. Her continuous support, trust, and encouragement to seek interdisciplinary collaborations and pursue divergent problems from various fields made this dissertation what it is now. Thank you!

My sincere gratitude goes out to the two external reviewers, Ofer Arieli and Jan Broersen, who devoted such a valuable portion of their time to scrutinizing this dissertation. Their careful comments helped to shape this thesis into its present form.

Throughout this PhD, I was lucky to meet many wonderful people who introduced me to the various exciting research topics that ended up in this thesis. My profound thanks go out to my co-advisor Stefan Woltran for introducing me to the fascinating field of formal argumentation; to Elisa Freschi for her inspiring approach to the intriguing world of Mīmāṃsā; to Sanjay Modgil for warmly welcoming me to London and exciting my evergrowing interest in dialogue models; to Jean Wagemans for exposing me to the richness and complexity of argumentation theory; to Christian Straßer for enthusiastically receiving me in Bochum and demonstrating the great potential of combining proof theory and formal argumentation; and to Leon van der Torre for serving as an informal advisor during his two years, on and off, sabbatical at TU Wien during which I was fortunate to get a glimpse of his profound knowledge of everything deontic logic.

A PhD makes for an exciting, adventurous, and trying life. It becomes great fun through the people with whom one is lucky to engage on a daily basis. It is unbelievable how many people one can meet over a span of five years. This half a decade at TU Wien gave me the pleasure of sharing many scientific, musical, and personal moments with so many lovely people: Alexandra, Anela, Andre, Anna R., Anna L., Björn, Clara, David, Dim, Dominik, Emery, Francesco, Francesca, Iana, Ilya, Ilina, Jan, Josephin, Matteo, Maya, Markus, Martin, Medina, Niklas, Rafael, Revantha, Robert, Roman, Sanja, Tim, Timo, Tiziano, Tobias, Xavier, and all of you who joined our TU jam sessions over the years. A special thanks to Anna Prianichnikova and Beatrix Buhl for always being ready for organizational matters and to Ilya and Tiziano for proofreading parts of this thesis.

This last year could not have been such a joyful experience without all the music made in Wien! Thanks to all my musician friends for recharging my battery every time we played.

I am more than grateful for my parents' unconditional support over the last 33 years. Thank you, Jet and Bert.

Els, thank you for being on this journey together. Your support means the world to me. Yaay!

Abstract

In this thesis, we employ logic to increase our understanding of normative reasoning. We do this by including agents in our formal analysis. Norms are inextricably linked to agents: they provide reasons to act and influence how we shape our world. Nevertheless, agentive aspects are often abstracted away, yielding oversimplified formalisms and understudied themes. Furthermore, recent developments in Artificial Intelligence (AI) have created novel challenges for the logical study of normative reasoning. This thesis addresses several of these topics by assigning a pivotal position to agents. The conducted research is interdisciplinary, drawing from methods in philosophy, logic, and AI.

The thesis comprises three parts: (I) agency, (II) action, and (III) argumentation. In the first two parts, we address normative reasoning by reasoning about agents. These parts belong to well-established modal logic approaches in deontic logic. In the third part, we employ methods from AI—in particular, formal argumentation and nonmonotonic logic—to make formal normative reasoning more accessible to agents.

In Part I, we investigate how obligations impact the choices of agents. To do so, we adopt and extend the agency logic of ‘Seeing To It That’ (STIT). We formally investigate the limits of contrary-to-duty reasoning when reasoning about choices and obligations over time. Furthermore, we conduct a comprehensive logical study of the principle of Ought implies Can. We investigate ten interpretations of the principle and its relation to other normative reasoning principles.

In Part II, we study ways in which obligations and prohibitions promote the actions performed by agents. We focus on instrumentality statements, which express actions as instruments for attaining ends. We develop a deontic action logic in which we analyze, compare, and assess obligations and prohibitions about instruments. Furthermore, we apply our formalism to an ancient theory in Sanskrit philosophy that reduces obligations and prohibitions to instrumentality statements.

In Part III, we investigate explanations in the context of defeasible normative reasoning. We call these deontic explanations. We develop a sequent-style proof-theoretic approach tailored to generating explanatory arguments and show how these arguments can be used in formal argumentation to create explanations. Furthermore, we develop a general, nonmonotonic proof-theoretic formalism that incorporates argumentative concepts like attack and defense, and extend it to defeasible normative reasoning.

Kurzfassung

In dieser Diplomarbeit verwenden wir Logik, um unser Verständnis des normativen Denkens[1] zu erweitern, indem wir Agenten in unseren formalen Analysen eine zentrale Position zuweisen. Normen sind untrennbar mit Agenten verbunden: Sie geben Gründe für Handlungen und beeinflussen, wie wir unsere Welt gestalten. Aspekte von Agenten werden jedoch oft abstrahiert, was zu vereinfachten Formalismen und unteruntersuchten Themen führt. Außerdem haben jüngste Entwicklungen in der Künstlichen Intelligenz (KI) neue Herausforderungen für die logische Untersuchung des normativen Denkens geschaffen. Diese Arbeit behandelt mehrere dieser Themen. Die resultierende Forschung ist interdisziplinär und nutzt Methoden aus Philosophie, Logik, und KI.

Die Arbeit besteht aus drei Teilen: (I) Agentialität (Agency), (II) Handlung (Action), und (III) Argumentation. Die ersten zwei Teilen gehören zu den modallogische Ansätzen der deontischen Logik und handeln vom Denken *über* Agenten im Kontext von Normen. Im dritten Teil nutzen wir KI-Methoden—insbesondere formale Argumentation und nichtmonotone Logik—um formales normatives Denken *für* Agenten zugänglicher zu machen.

In Teil I untersuchen wir wie Verpflichtungen die Entscheidungen von Agenten beeinflussen. Hierfür übernehmen wir die Logik von 'Seeing To It That' (STIT) und erweitern diese. Wir untersuchen formal die Grenzen des contrary-to-duty (das heißt "entgegen der Pflicht") Schließens in einem expliziten zeitlichen Kontext. Darüber hinaus führen wir eine umfassende logische Studie des Prinzips von Ought implies Can (das heißt "Sollen impliziert Können") durch. Wir untersuchen zehn Interpretationen des Prinzips und die Beziehung zu anderen Prinzipien des normativen Denkens.

In Teil II untersuchen wir wie Verpflichtungen und Verbote die expliziten Handlungen von Agenten fördern. Hierbei konzentrieren wir uns auf Urteile der Instrumentalität, die Handlungen als Instrumente zur Erreichung von Zielen ausdrücken. Wir entwickeln eine deontische Handlungslogik, in der wir Verpflichtungen und Verbote in Bezug auf Instrumente analysieren, vergleichen und bewerten. Außerdem wenden wir unseren Formalismus auf eine antike Theorie in der Sanskrit-Philosophie, die Verpflichtungen und Verbote auf Urteile der Instrumentalität reduziert, an.

[1] Im Kontext der Kurzfassung kann normatives Denken (*normative reasoning*) als das Ziehen von Schlussfolgerungen auf Grundlage von Normen, Pflichten, Erlaubnissen und Verboten definiert werden.

In Teil III untersuchen wir Erklärungen im Kontext des widerlegbaren normativen Denkens. Wir nennen diese deontische Erklärungen. Wir verwenden einen beweistheoretischen Ansatz und entwickeln einen Sequenzenkalkül, der darauf abzielt, erklärende Argumente zu generieren. Darüber hinaus, demonstrieren wir wie diese Argumente in formaler Argumentation verwendet werden können, um Erklärungen zu erstellen. Anschließend entwickeln wir einen allgemeinen, nichtmonotonen beweistheoretischen Formalismus, der argumentative Konzepte wie Attacke und Abwehr in der Sprache des Kalküls integriert. Wir wenden diesen Formalismus auf widerlegbares normatives Denken an.

Table of Contents

CHAPTER 1

Introduction

This thesis is about normative reasoning. Normative reasoning involves obligations, prohibitions, permissions, rights, values, and norms. These notions are everywhere. They make up laws, ethics, morals, business protocols, games, and social customs. What is more, they are directed to us agents. They influence our everyday decision-making and shape our lives. The obligation "promises must be kept" might cause us to hurry up and be on time for an appointment. Whereas the prohibition "it is not allowed to drive through a red light" affects whether or not we stop and wait for a traffic light to turn green. Moreover, we often find ourselves in situations where various norms conflict. In such cases, we must resolve these conflicts and decide which norms to give precedence. For instance, I might violate my promise to be on time because I decide to wait for a red light, whereas someone else may actually decide to do the opposite. Norms and normative concepts motivate, guide, redirect, and inspire the way we behave and how we form our world: they provide reasons to act.

In particular, this thesis provides a logical analysis of normative reasoning in the context of agents. Agents make choices, perform actions, and exercise abilities. They reason practically about attaining ends, plan short-term and long-term, and comply with and violate norms. Furthermore, agents may not always understand why certain norms apply, and obligations hold. Sometimes they may even disagree with the reasons given. We address several of these aspects by dividing this thesis into three parts:

I. Agency: obligations as restricted by the choices and abilities of agents;

II. Action: the interaction between normative concepts, actions, and instruments;

III. Argumentation: explaining to agents why certain norms do or do not apply.

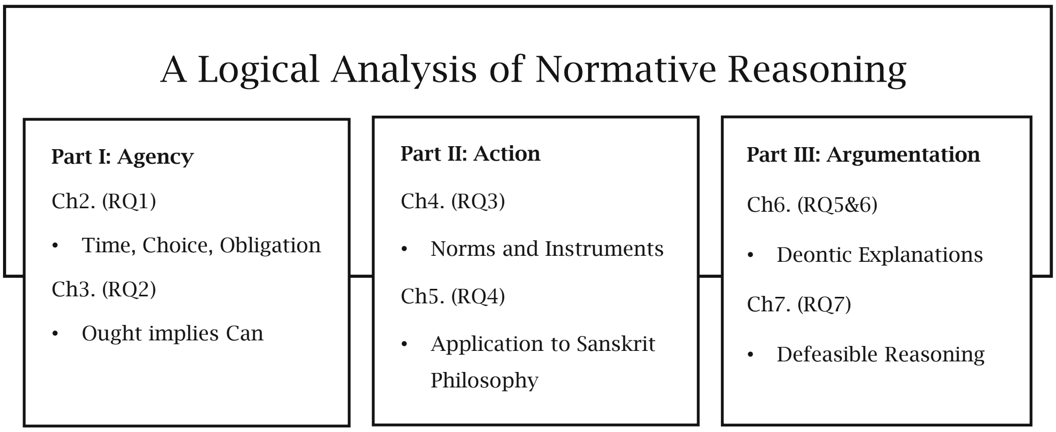


Figure 1.1: The general structure of the thesis.

Each of these topics represents fundamental challenges and questions that belong to the study of normative reasoning. In the following section, these topics are introduced and discussed in detail.

1.1 Guiding Questions and Problems

Deontic Logic is the overarching term for the field of mathematical logic that deals with normative concepts such as obligation, permission, prohibition, and norms. Since the 1950s, a wide range of deontic logics has been introduced. These logics are most often modal logics (Blackburn et al., 2004), employing modal formulae such as $\mathcal{O}\varphi$ expressing "it is obligatory that φ" (where φ denotes a state of affairs). Such formulae represent *normative propositions* according to some underlying set of norms, i.e., a normative code (Hilpinen and McNamara, 2013). Norms are 'rules' and 'laws' (von Wright, 1963a), which are conditional statements from which obligations, prohibitions, and permissions, are inferred (Parent and van der Torre, 2013).

The term 'deontic' refers to the Greek word *δέον* signifying "that which is binding". Following Kelsen (1991), a norm's object of binding is agentive behavior. The current view includes the behavior of humans as well as of artificial agents (Floridi and Sanders, 2004). Even though most approaches in deontic logic abstract away this 'object', many developments in the field have been guided by the conviction that agency and action are pivotal components of normative reasoning (Castañeda, 1972; Meyer, 1988; von Wright, 1968). In fact, the prominence of agents and acts can be traced back to the introduction of deontic logic by von Wright (1951).

Research on the formal aspects of normative reasoning and agents is vast but various problems remain un(satisfactorily)-addressed. What is more, novel problems arise due to recent developments in Artificial Intelligence (AI) and Normative Multi-agent Systems

(NorMAS). We provide a systematic treatment of several challenges from the field by dividing this thesis into three parts: *agency, action,* and *argumentation.*

The first two parts correspond to two well-developed approaches to including agents in the logical analysis of normative reasoning (Broersen et al., 2013). The first part focuses on *agency* and adopts the view that obligations (and other normative concepts) reciprocally impact agents' choices and what is brought about by them. The second part concentrates on *action* as performed by agents. Here, we take obligations (and other normative concepts) to promote and demote the actions agents can undertake to attain their ends. The third part of this thesis deals with *argumentation* as a means of characterizing and explaining conflicts between norms. The relatively new field of formal argumentation studies arguments and their relations, and is particularly suitable for the representation of conflicts.

In what follows, we elaborate on the three parts. Each part is divided into two chapters. We introduce the preliminary background and pose our general research questions along the way. In the remainder of this thesis, these questions are further specified and concrete objectives are presented. The general structure of the thesis is presented in Figure 1.1.

1.1.1 Part I: Agency and Normative Reasoning

The first part of the thesis deals with *agency*, namely, the idea that agents are capable of making *choices* and, consequently, influencing and changing the state of the world. Agency, taken in this sense, adopts an indeterministic worldview (Hilpinen, 1997): time progresses but may evolve in various ways. The future is not fully determined, and by choosing and acting, agents provide an essential contribution to delimiting possible courses of events.[1] Likewise, normative concepts such as obligations influence the choices made by agents and, indirectly, the state of the world. For instance, the fact that I ought to hand in my thesis next week may influence whether I go to a concert tonight.

Chapter 2. The interaction between time and obligation is one of the central research themes in deontic logic (Broersen and Torre, 2011). Although most deontic formalisms do not employ explicit temporal operators (Broersen et al., 2013), temporal deontic logics have been thoroughly investigated (Broersen et al., 2004; Dignum and Kuiper, 1997; van Eck, 1982; Prakken and Sergot, 1996; Thomason, 1981). Arguably the most prevailing logic of agency covering choice, time, and deontic concepts is the logic of '<u>S</u>eeing <u>T</u>o <u>I</u>t <u>T</u>hat' (STIT, for short). It was initially developed by Belnap and Perloff (1988) to semantically model agents' choices in indeterministic time. The STIT formalism is a modal logic, with its primary modality $[i]$ expressing that "agent i sees to it that" (some state of affairs hold). Over the past decades, a vast body of research has been developed around STIT. The formalism has seen a wide range of applications, covering, among

[1] In AI, the intelligent autonomous agent metaphor focuses primarily on the individual agent perspective, whereas for NorMAS, the focal point is that of multi-agent interaction (Verhagen et al., 2018). We deal predominantly with individual agency. We do not consider groups of agents (Herzig and Schwarzentruber, 2008) or agents as creators, modifiers, and enforcers of norms (Boella et al., 2008).

others, legal reasoning (Armgardt et al., 2018; Lorini and Sartor, 2014; Lorini and Sartor, 2015), epistemic reasoning (Broersen, 2008; Broersen, 2011a; Abarca and Broersen, 2019), and reasoning for autonomous vehicles (Arkoudas et al., 2005; Shea-Blymyer and Abbas, 2021).

Deontic extensions of the STIT framework were discussed since the beginning of STIT (Belnap and Perloff, 1988; Belnap, 1991; Bartha, 1993), but its first extensive investigation was provided by Horty (2001). In particular, Horty argues that temporal settings provide good reasons for adopting more refined notions of agentive and conditional obligations. These observations lead Horty to develop his influential deontic theory of *dominance act utilitarianism* in STIT. However, these observations are grounded in the interaction between deontic modalities and the *implicit* underlying semantic framework of indeterministic time. Surprisingly, a corresponding logic of explicit temporal deontic reasoning in STIT has not yet been developed. We set out to do this. The first research question pursued in this thesis is phrased accordingly:

Research question 1. *How can we model reasoning about obligation and choice in an explicitly indeterministic temporal setting? What are the logical and philosophical consequences of such a model for normative reasoning in the context of* STIT*?*

The first step towards answering the above questions is the development of a Temporal Deontic STIT logic. Although the semantic interpretation of indeterministic time—represented through branching-time frames (Prior, 1967; Thomason, 1981)—was present from the outset in STIT (Belnap and Perloff, 1988), the first sound and complete axiomatization of temporal STIT logic with branching-time frames was provided only a decade ago by Lorini (2013).[2] Likewise, the first technical results concerning Horty's deontic STIT logic were obtained by Murakami (2005), who proved the proposed theory of dominance ought sound, complete, and decidable. As a consequence of addressing the above research question, we fill a longstanding literature gap by providing the first sound and complete axiomatization of temporal deontic STIT logic.[3] What is more, the obtained framework enables us to analyze the logical consequences of normative reasoning in an explicit temporal setting. In particular, it enables us to reassess some of Horty's (2001) observations formally.

Chapter 3. The choices that agents make depend on their abilities. Obligations and prohibitions often depend on the abilities of agents too. In fact, one of the most ubiquitous principles governing normative codes and ethical systems is the metaethical principle called *Ought implies Can* (OiC). Intuitively, the principle states that "each obligation presupposes a possibility of fulfilling it" (Hintikka, 1970, p.83). OiC has a

[2] See also (Armgardt et al., 2018; Broersen et al., 2006; Broersen, 2008; van Berkel and Lyon, 2019b; Ciuni and Lorini, 2018; Wansing, 2006) for other temporal characterizations of STIT.

[3] Alternative deontic temporal STIT logics are given by Broersen (2008) and Lorini (2013). Both take deontic concepts as *defined* by using violation constants in the spirit of Anderson (1958) (the former uses the implicitly temporal XSTIT choice operator). See page 8 for an introduction to Anderson's approach.

long history within philosophy and has been traced back to Aristotle (*The Nicomachean Ethics*, translated by Ameriks and Clarke, 2000), ancient Roman law (Vranas, 2007), and Immanuel Kant (*Critique of Pure Reason*, translated by Guyer and Wood, 1998). Over the past decades, OiC became a topic of investigation in its own right (Copp, 2017; Kohl, 2015; McConnell, 1989; Stocker, 1971). The result is a vast body of literature on the topic containing a variety of interpretations of the principle (van Ackeren and Kühler, 2015; Vranas, 2007). OiC is not uncontroversial and, despite its importance, there is no clear consensus on its interpretation, let alone its implications. Determining the right interpretation of OiC is crucial for normative systems that adopt it since it influences the degree to which an agent can be burdened with and relieved from duties (Dahl, 1974; McConnell, 1989). For instance, can I be obliged to take my bike to work this morning if my bike was stolen? Moreover, since OiC is part of various ethical and (ancient) legal systems, it becomes all the more important to understand OiC and its various readings better. Formal models provide an effective way of increasing our knowledge in this respect.

Indeed, in one way or another, OiC is already a principle of most deontic logics (Hilpinen and McNamara, 2013). What is more, OiC is said to be one of those deontic logic properties commonly taken as 'undisputed' in the field (van der Torre, 1997). Surprisingly, there is a severe discrepancy between the philosophical and logical approaches to OiC. For instance, in philosophy, the principle is predominantly taken as agentive: "what ought to be done, can be done"; cf. (van Ackeren and Kühler, 2015). Nonetheless, in deontic logic, the principle is commonly taken as impersonal: "what ought to be, is possible"; cf. (Hilpinen and McNamara, 2013). Thus, there is a significant gap between the formal treatment of OiC and its philosophical counterpart that it aims to model. The second research question concerns the formal analysis of Ought implies Can:

Research question 2. *What are the logical relations between the various readings of Ought implies Can encountered in philosophy? What are the consequences of these readings for formal normative reasoning?*

In answering these questions, we find that the level of abstraction adopted in most deontic logics keeps us from capturing essential nuances and refinements necessary for an adequate analysis of Ought implies Can. We address the above questions within the formalism of deontic STIT and develop a class of deontic STIT logics with which we analyze ten philosophical readings of OiC. Furthermore, we employ these logics to formally investigate how other pivotal principles relate to OiC (see page 16 for an overview). The primary motivation for adopting STIT is that STIT provides a formal language conducive to modeling various refined agentive concepts, such as 'ability', 'refrainability', and 'deliberative choice' (Belnap et al., 2001). Alternative formalisms, in this respect, are the logic of ability by Brown (1988) and the logic of 'bringing it about that' by Elgesem (1997). In contrast to these alternatives, the STIT formalism is highly modular, and deontic extensions of STIT are well-developed (see research question 1).

1.1.2 Part II: Action and Normative Reasoning

The second part of this thesis deals with the performance of *actions* in relation to normative concepts such as obligation and prohibition. Whereas Part I deals with the choices available to agents and the outcomes thereof, Part II treats explicit actions as first-class citizens. Here too, we take action to assume indeterministic time at its base: at each given moment in time, an agent can perform various actions which may or may not ensure an envisioned outcome. This view takes every action to be associated with a change, namely, a transition between two states (von Wright, 1963a; Hilpinen, 1997). It supports the decision to consider actions as syntactically different from propositions describing states of affairs. A common approach to upholding this distinction is by using modalities for actions, such as in Propositional Dynamic Logic (PDL, for short), where an action modality $[\delta]$ is interpreted as "the performance of action δ ensures that" (some state of affairs holds) (Åqvist, 1974; Fischer and Ladner, 1979).[4] Such logics are also referred to as dynamic action logics. The formalism was first adapted to the context of normative reasoning by Meyer (1988) and continues to receive attention to the present day, e.g., see (Giordani and Canavotto, 2016; Giordani and Pascucci, 2022; Hughes et al., 2007).

Chapter 4. Normative codes prescribe (or prohibit, for that matter) certain states of affairs and the performance of particular actions. *Instrumentality statements*—or means-end relations—fulfill an essential role in this respect. They describe how actions serve as instruments (means) for the attainment of desired outcomes (ends) (Condoravdi and Lauer, 2016; Hughes et al., 2007; von Wright, 1972b). These statements are central to practical reasoning and deliberation and guide an agent towards achieving her goals (Bratman, 1981; Hare, 1971; von Wright, 1972a). Furthermore, means-end reasoning plays a central role in Belief-Desire-Intention (BDI) logics (Rao and Georgeff, 1995) and related multi-agent systems (Dastani, 2008), where means are considered as plans that stipulate a sequence of (sub)actions needed to attain a given goal (Rao and Georgeff, 1998). Norms play an important role in this respect: they can prescribe or forbid the attainment of certain ends on the one hand and the performance of particular actions on the other. In deontic logic, this twofold role assigned to norms is well-studied and, in the case of obligations, it is referred to as the dichotomy between *ought-to-be* and *ought-to-do* (Castañeda, 1972). The dichotomy is an important challenge for deontic logic and NorMAS, where the main question is whether the latter can be reduced to the former (Pigozzi and van der Torre, 2018).

There is, however, another role that norms can play in relation to instrumentality statements, and that is when such statements form the content of obligations and prohibitions. To see this, consider the following example: "It is prohibited to use nonpublic information as an instrument to acquire financial profit on the stock market". This prohibition is known as the law on 'insider trading'. Notice that it is neither

[4] St. Anselm (1033 – 1109) is said to be the first to investigate the logical structure of action (Segerberg, 1992). Hilpinen (1997) provides a detailed history of action logic.

prohibited to use nonpublic information nor is it prohibited to acquire financial profit on the stock market. Only as a means to attain financial profit, using such information is forbidden. Prohibitions of the form expressed above articulate which actions may not be employed as instruments for achieving particular goals. We call obligations and prohibitions belonging to this category *norms of instrumentality*. Despite the ubiquity of normative constraints on instrumentality in legal, social, and ethical systems—think of protocols, rules of games, and fairness constraints—an investigation of their philosophical and logical ramifications in logic is absent. This thesis sets out to fill this knowledge gap, providing the foundations for formal reasoning with norms of instrumentality.

Research question 3. *How can we formally represent obligations and prohibitions about instrumentality statements? Moreover, how do they relate to the dichotomy between ought-to-be and ought-to-do statements?*

We address these questions by providing a modal deontic action logic based on Wright's theory of agency and instruments (1963b,1972b). Our approach deviates from the traditional approach to dynamic action logic. Namely, we propose a *reduction of action modalities* to alethic formulae containing action constants functioning as witnesses, i.e., "action δ is performed by agent i when the next moment witnesses the successful performance of δ by agent i". The use of action witnesses as constants preserves the critical view of actions as distinct, first-class citizens in the formal language. The resulting language accommodates formalizations of various notions of instrumentality. We point out that this thesis is not concerned with instrumentality in planning and BDI logics (Meyer et al., 2015). Instead, we investigate norms *about* instrumentality relations and the logical properties of the obligations and prohibitions that result from them.

Chapter 5. As an application of instrumentality in the context of normative reasoning, we investigate the deontic theory of the south Asian Sanskrit philosopher Maṇḍana miśra (8CE), Maṇḍana, for short (Freschi, 2010). Maṇḍana belongs to the school of Mīmāṃsā, which is one of the most important schools of Indian philosophy with a long and rich history of investigating normative reasoning (Ciabattoni et al., 2015). The school—active for over two millennia—focuses on the exegesis and systematization of the prescriptive parts of the Vedas, the sacred texts of what is now called Hinduism. Mīmāṃsā authors invested much effort in rationally interpreting Vedic commands and resolving conflicts. The result is a vast body of rigorously structured theories of normative reasoning. Maṇḍana's deontic theory is unique because it contains a *deontic reduction*: i.e., a uniform reduction of all Vedic commands to purely descriptive statements about actions *instrumental* to desirable and undesirable outcomes. For Maṇḍana, a command such as "If one desires rain, one should perform the Kārīri ritual" is reduced to the descriptive statement "The Kārīri is an instrument for attaining rain".

Due to their highly systematic nature, Mīmāṃsā theories continue to be important to numerous fields, including, among others, Indian jurisprudence (McCrea, 2010). However, various Mīmāṃsā doctrines are still unexplored or misunderstood despite their undeniable importance. In particular, how Maṇḍana's deontic reduction relates to normative reasoning principles prevailing in the common Mīmāṃsā tradition remains to be determined. Logic provides an effective formal tool for a rigorous analysis of these doctrines (Ciabattoni et al., 2015; Freschi et al., 2017). In fact, there are some striking similarities between Maṇḍana's reduction and what is known as Anderson's reduction (Anderson, 1958) in deontic logic. An Andersonian reduction reduces deontic statements of the form "It is obligatory to stop for a red light" to statements of the form "Not stopping for a red light necessary leads to a violation" (Anderson and Moore, 1957; Castañeda, 1972). Deontic action logics that adopt the Andersonian reduction deal considerably well with challenging benchmark examples from the literature (Meyer, 1988). Such examples are referred to as deontic puzzles or paradoxes (Hilpinen and McNamara, 2013) and are discussed on page 13. Accordingly, we investigate the following questions:

Research question 4. *How can we formalize Maṇḍana's deontic theory, and how does the theory relate to the common Mīmāṃsā tradition? How does the resulting logic deal with contemporary deontic paradoxes?*

To address the above questions, we adopt the formal language developed in Chapter 4 and tailor the corresponding logic to Maṇḍana's theory of normative reasoning. The purpose of this Sanskrit application is twofold: First, we show that logic can deepen our understanding of Maṇḍana's deontic reduction and its position in the Mīmāṃsā tradition. Second, we formalize Maṇḍana's theory to demonstrate the advantages of incorporating instrumentality relations in the formal analysis of normative reasoning, e.g., in relation to dealing with deontic paradoxes.

1.1.3 Part III: Argumentation and Normative Reasoning

The first two parts of this thesis belong to the modal logic tradition in deontic logic. In contrast, Part III employs AI methods from the field of formal argumentation and defeasible reasoning to address the novel challenge of generating *explanations* in the context of normative reasoning. Some preliminaries are required.

Defeasibility. First, Part III deals with *defeasible* normative reasoning. We reason defeasibly when we draw conclusions due to the absence of information to the contrary (Reiter, 1980), when our reasoning is rationally compelling but not necessarily deductively valid (Koons, 2022), or when we jump to conclusions on the basis of normality, typicality, and probability (Straßer, 2014). In all these readings, the premises justify the conclusion even though additional information may force one to retract the conclusion later on (Pollock, 1987). Most of our daily life reasoning is defeasible (Toulmin, 1958) and, due to the presence of conflicts, violations, exceptions, and priorities, normative reasoning is inherently defeasible too (Nute, 1997). Formal systems of defeasible reasoning emerged in the 1980s due to rapid developments in AI[5] and logical systems of defeasible normative

[5] In particular, see the seminal *Special Issue on Non-Monotonic Logic* (Bobrow, 1980) of the Artificial Intelligence journal. For a historic overview, we refer to the work of Koons (2022).

reasoning were introduced soon after (Horty, 1997; Makinson and van der Torre, 2001; Governatori and Rotolo, 2006). The central characteristic of formalisms of defeasible reasoning is that they are *nonmonotonic*. That is, they do not satisfy the property of monotonicity, which ensures that inference of a conclusion from a set of premises is robust under expansions of those premises. In other words, nonmonotonic formalisms allow for the retraction of a conclusion in the light of additional information (Straßer, 2014).

Argumentation. Formal argumentation provides a uniform theory of nonmonotonic reasoning. Namely, many nonmonotonic logics can be represented in formal argumentation yielding the same inference relation as the characterized logic. See the work of Arieli et al. (2021) and Straßer (2014) for an overview of these results.[6] The central concept of this field is that of an instantiated argumentation framework, introduced by Dung (1995). It consists of a set of arguments—where arguments comprise a claim and a collection of reasons in support of it—together with an attack relation that defines conflicts between these arguments. Furthermore, semantic extensions are identified as sets of justified arguments collectively defendable against counterarguments. The idea of defeasibility is then captured in terms of *counterarguments* attacking an initial argument. For instance, I may argue that Franz can sing because Franz is a bird. An argument I may need to retract after you counterargue that Franz is an ostrich, and ostriches cannot sing.

Chapter 6. By providing argumentative characterizations of nonmonotonic deontic logics, we can harness existing methods from the field of formal argumentation and apply them to the context of normative reasoning. The most promising and well-studied formalisms in this respect is that of *Input/Output logic* (I/O logic, for short) (Makinson and van der Torre, 2001). In brief, I/O logics model normative systems that stipulate how to contextually detach obligations and permissions from a normative code (Parent and van der Torre, 2013). In particular, nonmonotonic I/O logics (Makinson and van der Torre, 2001; Parent, 2011) have been employed to defeasibly reason with deontic conflicts, norm violations, and exceptions. Some first results concerning argumentative characterizations of normative reasoning and the I/O formalism have been obtained (Straßer and Arieli, 2015; Liao et al., 2018; Straßer and Pardo, 2021).[7] However, much work needs to be done. For instance, these approaches consider only fragments of the standard I/O systems, employing languages restricted to literals. What is more, these approaches are not suitable for explanatory purposes. For instance, Liao et al. (2018) and Straßer and Pardo (2021) take arguments to consist only of norms and not of inferences,

[6] Prakken (2018) identifies two views on formal argumentation: argumentation as inference and argumentation as dialogue. We adopt the view of argumentation as inference. The literature on formal dialogues is vast: ranging from inquiry, information-seeking, and persuasion dialogues of argumentation, to dialogues of practical deliberation (Black and Hunter, 2007; McBurney and Parsons, 2009). A number of these works provide for dialogical generalizations of argumentation as inference, where two agents discuss the acceptability of a given argument, e.g., (Prakken, 2005).

[7] See the work of Dong et al. (2020) and Governatori et al. (2018) for argumentative characterizations of other deontic logics.

and for Straßer and Arieli (2015), arguments may contain irrelevant information. Part III continues this research program.

Research question 5. *How can we provide a modular logical formalism that yields argumentative characterizations of a large class of nonmonotonic Input/Output logics?*

Once such a characterization is obtained, we can start applying existing methods from formal argumentation to I/O reasoning. We are interested in one such application: *explanations*. The use of formal argumentation for explanatory purposes is promising and the field is rapidly expanding (Borg and Bex, 2021; Čyras et al., 2021).

In the context of agents and normative reasoning, explanations are critically important. In order to motivate compliant behavior, an agent must understand why she is required to behave in a specific way (particularly if she disagrees with her alleged duty). In this respect, it does not suffice for the agent to know that an obligation holds: she must know *why* it holds. Especially in view of normative conflicts, answers to such why questions become crucial. Consider the question "why am I permitted to take over on the left, despite my obligation to drive on the right?". A satisfactory answer not only explains that I am permitted but also why the other obligation does not hold. In the above example, the permission can be an exception to the obligation, thus making the latter inapplicable in the context of taking over other vehicles. We call answers to such questions *deontic explanations*[8]. Such explanations not only improve an agent's understanding of norms but also provide reasons that motivate the agent's appropriate conduct. Most deontic logics only show that some obligation holds, and deontic logic has not yet been investigated with the aim of generating explanations. In fact, explaining normative reasoning is identified by Peirera et al. (2017) as one of the three challenges for formal argumentation approaches to NorMAS.

Research question 6. *How can we accommodate deontic explanations in the developed argumentative characterizations of the various Input/Output logics?*

Enhancing the explainability of I/O reasoning is an attempt to optimize the existing expressivity of the I/O formalism. We can think of research question 5 as laying the foundation for answering research question 6. We address research questions 5 and 6 by developing *sequent-style* proof calculi that generate explanatory I/O arguments. The sequent calculus formalism—originating in the work of Gentzen (1934)—defines proof systems in terms of sets of *rules*. One of the principal characteristics of sequent systems (compared to Hilbert-style axiomatic proof systems) is the rule-based approach to constructing proofs as trees: the leaves of a tree are either trivial logical truths or assumptions, branches are the result of rule applications, and the tree's root is the conclusion (Negri et al., 2008).[9] Over the past decades, the sequent framework has been

[8] The term was suggested by Agata Ciabattoni, Christian Straßer, and Leon van der Torre.

[9] Sequent calculi can be shown to possess the property of *analyticity*, which expresses that any derivable formula is derivable with a proof solely consisting of subformulae of the formula in question.

extended to cover a wide range of logics and formalisms, including modal logics (Negri, 2005). We are primarily interested in developing classes of sequent systems that generate logical arguments that show a strong correspondence with formal argumentation, e.g., see the work of Arieli and Straßer (2015).

Chapter 7. One of the main contributions of Chapter 6 is the development of a class of sequent-style proof systems characterizing monotonic normative reasoning in the I/O formalism. The main technical result of that chapter is that a large class of nonmonotonic I/O logics can be argumentatively characterized through argumentation frameworks instantiated with arguments generated by these proof systems. An immediate question is whether the developed proof systems can be modularly extended with rules that directly capture defeasible normative reasoning. In Chapter 7, we address this question by pursuing a *more general aim.* For this, we make use of the following observation: An essential feature of defeasible reasoning (and nonmonotonic logics) is that previous inferences can be retracted in the light of novel information. To illustrate, one may find that an initial obligation "you ought to drive on the right side of the road", must be revised in the context of an exceptional circumstance, e.g., when overtaking another vehicle. Thus, we say, in the context of defeasible reasoning, the *status* of a formula as a logical conclusion may have to be *revised* (several times). The research question pursued in this final chapter is formulated accordingly:

Research question 7. *How can we integrate status revision considerations of defeasible reasoning into the object level of sequent-style proof systems? Can we show these proof systems to yield a nonmonotonic inference relation?*

In formal argumentation, the process of revision is intuitively represented in terms of attack and defense. Our primary objective is to integrate the central concepts of *revision, attack,* and *defense* into a proof-theoretic approach to nonmonotonic logic. The expression of attack and defeat in sequent-style proof systems has been extensively investigated, e.g., by Arieli and Straßer (2015; 2019). The main difference with our objective is that existing approaches leave both revision and nonmonotonic inference for the meta-analysis of the proof system. We set out to incorporate these features on the level of the proof. The result is a novel proof-theoretic approach for nonmonotonic reasoning with conflicting information in which revision procedures are fully integrated on the object level of proofs. We will demonstrate that nonmonotonic inference in our calculi strongly relates to various types of inference in formal argumentation. Returning to our initial aim, we will leverage these results to enhance the calculi from Chapter 6 to obtain a class of nonmonotonic proof systems for normative reasoning. A direct advantage of the approach is that we can express normative conflicts in the object language of proofs by employing the integrated notions of attack and defense.

That is, one can construct proofs by merely decomposing a formula, making it an effective tool for proof search. Analyticity is also useful for determining other properties, such as consistency of a logic.

Situating the Thesis

The logical analysis of normative reasoning is an interdisciplinary research field. The results of this thesis are specifically relevant to philosophical logic and Knowledge Representation and Reasoning (KR). As a subfield of philosophy and logic, philosophical logic deals with applying logical methods to problems in philosophy. The main aim is to enhance our understanding of these problems through mathematical analysis. Think of problems concerning properties of time, knowledge, and norms. The use of modal logic is the predominant approach in this field. Parts I and II contribute to this field. As a subfield of AI, KR deals, among others, with the formal representation of knowledge for reasoning tasks. We find various modifications of philosophical logics employed in the context of KR. Think of temporal, epistemic, and deontic logics. In particular, the I/O formalism is highly suitable for defeasible normative reasoning tasks, representing normative systems as knowledge bases. Part III primarily contributes to KR.

1.2 Methodology

The field of deontic logic is not characterized by a principal methodology and the interdisciplinary research presented in this thesis involves methods from philosophy, logic, and AI. In this section, we reflect on various methods. Formal logics are, by definition, abstractions and this implies that design decisions must be made. Such choices largely depend on the reasons of formalization. In general, we can distinguish between theoretical and practical reasons. This thesis deals with both.

Concerning theoretical reasons, formal models provide mathematically precise means—i.e., analytic tools—for enhancing our understanding of concepts and reasoning with concepts. Formalization has the unique advantage of employing mathematical methods for evaluating such models with respect to, for instance, a set of formally specified properties the model should ideally satisfy (cf. page 16 below). Furthermore, mathematically precise languages facilitate the comparison of various logics modeling the same (or similar) phenomena. Part I and II involve such theoretical reasons.

Practical reasons are concerned with (defeasible) reasoning tasks, computability, explainability, and implementations. For instance, logics of normative reasoning can be harnessed for reasoning tasks that deal with conflicts, planning, and compliance checking. Another prominent method for practically assessing a developed logic is by determining its computational complexity.[10] Moreover, recent developments in AI show that formal models can be of specific use in generating explanations. Part III addresses practical reasons concerned with defeasible reasoning, conflict resolution, and explanation.

When developing a formal logic, the following two central questions must be addressed:

1. Which formalism should be employed in developing a formal system?

[10] Complexity considerations and implementation fall outside the scope of this thesis.

2. How can we assess the correctness or suitableness of a formal system with respect to the phenomenon it intends to model?

The above questions are strongly connected. The first question relates to the various types of formalisms available: For instance, one may adopt a propositional or a first-order language, a classical or an intuitionistic base, a modal approach, a nonmonotonic approach, and so on. Additionally, one may provide different semantic and proof-theoretic characterizations of the same logic. The choices made in this respect strongly depend on determining the right depth of abstraction. We address these considerations throughout this thesis in the respective chapters.

The second question concerns the evaluation of the obtained formalism. In the remainder of this section, we further elaborate on this question and discuss three methods often employed for *assessing* deontic logics: these concern deontic puzzles, metaethical principles, and philosophical foundations. We frequently refer to these methods throughout the rest of this thesis.

1.2.1 Examples and Deontic Puzzles

As in many other fields—such as formal argumentation (Caminada, 2004), linguistics (Condoravdi and Lauer, 2016; Kratzer, 1981), and ethics (Sverdlik, 1985)—examples and counterexamples play a central role for developments in deontic logic. Often, such examples are given in natural language and appeal to some common intuition. A benchmark example is a quintessential example that a formalism should be able to deal with. Failing to address such an example correctly does not necessarily imply refutation of the formalism at hand but, at minimum, forces one to reflect on whether to revise the formalism or whether the model constitutes an exception to the example.

The most prominent benchmark examples in deontic logic are *deontic puzzles* or *deontic paradoxes*. They are the driving force for defining and refining deontic systems (Hilpinen and McNamara, 2013). They highlight typical properties of normative reasoning and usually consist of the (un)derivability of certain formulae, counterintuitive to a given common-sense reading (van der Torre, 1997). We follow the suggestion of Hilpinen and McNamara (2013) and adopt the overarching term *puzzles* to denote these challenges. We distinguish between two types of puzzles: The first type emphasizes challenges of conditional obligations. The second type concerns unintuitive consequences of the logical properties of deontic systems. In various chapters of this thesis, we assess the developed formalisms in light of such puzzles. Here, we briefly recapitulate some of the most prominent ones and refer to Hilpinen and McNamara (2013) for an extensive overview.[11]

Remark 1.1. *Most of the problems indicated by the paradoxes pertain to Standard Deontic Logic (SDL), one of the oldest systems in deontic logic, e.g., see (Hilpinen and*

[11] This thesis investigates obligations, prohibitions, and norms. It does not deal with the study of permission, which can be taken as a proper subfield of deontic logic (Hansson, 2013). The study of permission comes with its own set of deontic puzzles (Hilpinen and McNamara, 2013).

McNamara, 2013). The language of SDL consists of the propositional connectives for negation ($\neg$), disjunction ($\vee$), conjunction ($\wedge$), and material implication ($\rightarrow$), together with the monadic modality $\mathcal{O}$ denoting "it is obligatory that" (some proposition holds). SDL is defined as the normal modal logic KD*, containing the deontic consistency axiom* $\neg(\mathcal{O}\varphi \wedge \mathcal{O}\neg\varphi)$*, often referred to as the* D*-axiom (Chellas, 1980). To facilitate discussion, some of the puzzles considered in this section contain, besides their natural language form, a formal representation in SDL.*

Puzzles of Conditional Obligations. Of all the challenges of conditional norms, contrary-to-duty (CTD) reasoning yields the most notorious challenge of them all (van der Torre, 1997; van der Torre and Tan, 1998). CTD reasoning deals with those obligations that hold whenever a violation ensues. There are many different CTD puzzles. Here, it suffices to discuss the original CTD paradox proposed by Chisholm (1963). The paradox pinpoints problems of conditional obligations and (deontic) detachment (Parent and van der Torre, 2017). It consists of the following four sentences:

(C1) Billy ought to go to the assistance of her neighbors.

(C2) If Billy goes to the assistance of her neighbors, she ought to tell them she is coming.

(C3) If Billy does not go to the assistance of her neighbors, she ought not to tell them that she is coming.

(C4) Billy is not going to the assistance of her neighbors.

The scenario expresses a CTD situation: (C1) is referred to as the unconditional initial obligation, (C2) is a compliant-with-duty obligation expressing an obligation conditional on the fulfillment of the initial obligation; and (C3) is a contrary-to-duty obligation expressing what ought to be in case a violation occurs. The last premise (C4) tells us that the agent is in a violation state contrary to her initial obligation. Following Hilpinen and McNamara (2013), "it is not at all as easy as it might seem to faithfully represent scenarios like those [...] and it proved to be a real shortcoming of the standard systems" (p.83) such as Standard Deontic Logic. The introduction of the paradox marks a turning point in deontic logic, initiating a thorough investigation of conditional obligations and the concept of violation.[12] Hilpinen and McNamara (2013) and Prakken and Sergot (1996) provide a (critical) overview of various solutions to the paradox. In Chapter 2, we discuss temporal CTD reasoning. In Chapter 5, we discuss an atemporal CTD scenario—i.e., the Gentle Murder Paradox (Forrester, 1984)—using an action-based deontic logic.

[12] This thesis treats obligations as monadic and defined conditional modalities. We do not investigate primitive dyadic operators. A dyadic obligation $\mathcal{O}(\varphi/\psi)$ expresses that "in the context ψ, it is obligatory that φ". Its antecedent enables one to single out what is obligatory in specific contexts (Parent, 2021). For this reason, dyadic deontic logics are suitable for reasoning about violation contexts (Chisholm, 1963) and differentiating between prima facie and all-things-considered obligations (Alchourrón, 1996).

In Chapter 6, we discuss deontic explanations of CTD scenarios in the context of the Input/Output formalism.

Puzzles of Logical Properties. One of the oldest deontic puzzles in the literature is Ross' Paradox (Ross, 1944). It consists in deriving from the obligation (R1) the counterintuitive obligation expressed in (R2):

(R1) It is obligatory that you mail the letter.

(R2) It is obligatory that you mail the letter or burn the letter.

In Standard Deontic Logic, the scenario is formalized as:

(r1) $\mathcal{O}(\texttt{mail})$

(r2) $\mathcal{O}(\texttt{mail} \vee \texttt{burn})$

The oddity of the puzzle is that (R1) implies the obligation (R2) which can be fulfilled by burning the letter—an act that violates the obligation in (R1). As Hilpinen and McNamara (2013) put it: "it remains odd to think I could plead partial mitigation in failing to mail the letter by burning it instead with 'Well, at least I fulfilled my obligation to mail or burn it' " (p.63). In Standard Deontic Logic, (r2) logically follows from (r1). Most normal modal logics suffer from this paradox due to the normality of the obligation modality (see Remark 1.1). A prominent solution to the paradox is to block the above logical inference by adopting a non-normal modal interpretation of the deontic modalities (Chellas, 1980). In Chapters 3 and 5, we discuss Ross' Paradox.

As indicated above, some puzzles may be solved by adopting a weaker—e.g., non-normal—modal interpretation of the obligation modality $\mathcal{O}$. However, there is a price to pay: one may lose some intuitively desirable inferential power of the logic in question. The Alternative Service Challenge (Van Fraassen, 1973) highlights this. The challenge consists in deriving from the following two commands:

(A1) You should fight in the army or perform alternative service.

(A2) You should not fight in the army.

The third command:

(A3) You should perform alternative service.

The inference of (A3) from the premises (A1) and (A2) is intuitively desirable. In fact, it is valid in Standard Deontic Logic where the scenario is represented as follows:

(a1) $\mathcal{O}(\texttt{fight} \vee \texttt{service})$

(a2) $\mathcal{O}(\neg\texttt{fight})$

(a3) $\mathcal{O}(\texttt{service})$

However, by adopting weaker modal logics—e.g., to solve other deontic puzzles—the inference of (a3) from (a1) and (a2) is often lost (Van Fraassen, 1973; Chellas, 1980; Horty, 1994). The challenge lies in solving some of the deontic puzzles while preserving certain desirable inferences. We discuss this challenge in the light of deontic dilemmas in Chapters 3 and 5.

1.2.2 Metaethical Principles

Another accepted way of assessing a developed logical systems is by comparing it with lists of principles the logic must ideally satisfy. Such principles are sometimes referred to as postulates or desiderata. Think of the rationality postulates in the field of formal argumentation (Caminada and Amgoud, 2007), the postulates of belief revision (Alchourrón et al., 1985), rationality principles for multi-agent systems (Dastani, 2008), and desirable properties of nonmonotonic inference (Arieli et al., 2022b; Straßer, 2014).

In normative reasoning, such criteria have been referred to as *metaethical principles* (McConnell, 1985), which are principles that *ideally* any ethical theory should satisfy. For instance, Ought implies Can is a metaethical principle. Metaethical principles can pinpoint flaws or weaknesses of such theories, giving rise to modifications and discussions.

Similarly, metaethical principles are central to the formal analysis of normative reasoning. We find applications of such principles since the early days of deontic logic (von Wright, 1951; Anderson and Moore, 1957). The ineptitude of particular formalisms to satisfy metaethical principles led to new developments in deontic logic. For example, the deficiency of normal modal logics to consistently represent deontic dilemmas led to the introduction and analysis of non-normal modal deontic logics (Chellas, 1980). In this thesis, we consider several such metaethical principles:

1. Ought implies Can: What an agent is obliged to do, an agent can do;
2. Deontic Consistency: Obligations are (individually) consistent;
3. Deontic Contingency: Obligations range over contingent states of affairs;
4. No deontic dilemmas: Obligations are jointly consistent;
5. Dilemmas are possible: Obligations can consistently require what is incompatible;
6. No Vacuous commands: Obligations are not about what trivially holds.

The above principles are phrased in terms of obligations, but one may likewise think of them in terms of prohibitions. It is not the case that failing to satisfy one of the above principles means the refutation of the proposed formal system. Some deontic theories may deliberately abstain from adopting some of these principles. Nevertheless, metaethical principles invite one to reflect critically on specific aspects of the formalism.

In Chapter 3, we thoroughly discuss the above list of principles and investigate their logical interdependencies. In Chapters 4 and 5, we investigate several such principles in the context of action and instrumentality.

1.2.3 Philosophical Foundations

Several sources can be used in order to formalize a particular phenomenon. One of which is philosophy. For deontic logics—but also agency logics and epistemic logics—one may build a formal system upon existing philosophical theories on the respective topics. The upshot of adopting this approach is that one's formal system is grounded in a thoroughly justified theory. Another advantage is that one can often lean on results that arose through a rich and long history of critical debates on the topic at hand. Moreover, such philosophical theories often provide a systematic study of principles, properties, and (counter)examples against which the formal systems can be evaluated.

To illustrate this point, we mention some influential philosophical theories in logic. Von Wright's theory of agency as developed in (von Wright, 1963a; von Wright, 1972b) has proven to be a rich source for developments in the logic of agency and action, e.g., (Åqvist, 2002; Segerberg, 2002). From the viewpoint of philosophy of law, Hohfeld serves as an essential source of inspiration (Glavaničová and Pascucci, 2021; Kanger, 1972; Markovich, 2020). Of recently, the ancient philosophical school of Mīmāṃsā—dealing with normative reasoning in the context of the Vedas—has been employed as a fruitful source for developing deontic systems, e.g., (Ciabattoni et al., 2015; Freschi et al., 2019; Lellmann et al., 2021; van Berkel et al., 2022a). The same applies to formal systems of Talmudic reasoning, e.g., (Abraham et al., 2011).[13]

Several chapters in this thesis are grounded in philosophical theories: Chapter 3 is based on an extensive survey of the philosophy of Ought implies Can, Chapter 4 is rooted in von Wright's philosophy of action and instrumentality, and Chapter 5 proposes a formalization of the deontic theory of one of the central authors of the school of Mīmāṃsā. Concerning the latter, we provide a more detailed discussion of our methodology in Chapter 5, which involves interpreting and translating Sanskrit texts.

[13] Another commonly recognized source of formalization is intuition. In fact, new fields may often largely depend on intuition as a driving forces behind developments due to the lack of a well-grounded theory. We refer to the work of Caminada (2004) for a discussion of intuition and the difference between intuition of lay people (*logica utens*) and those that arise through systematic study (*logica docens*).

1.3 Outline

Part I is devoted to analyzing normative reasoning in the context of agential choice.

In Chapter 2, we develop a sound and complete Temporal Deontic STIT logic (TDS). We show how the proposed semantics of TDS can be truth-preservingly transformed into the traditional deontic STIT semantics using utility functions. We demonstrate the limits of the traditional approach by providing an incompleteness result for explicit temporal contrary-to-duty reasoning in the context of deliberative agency.

Chapter 3 provides a comprehensive logical study of ten philosophical interpretations of Ought implies Can (OiC). We modify the deontic STIT formalism of Chapter 2 and develop a class of sound and complete deontic STIT logics (OS) axiomatizing these ten OiC interpretations. We employ the resulting logics to provide a formal taxonomy of the (in)dependencies of the various OiC readings. We then extend this class of STIT logics with other metaethical and normative reasoning principles and determine their relation to OiC.

Part II formally addresses instrumentality relations in the context of normative reasoning.

In Chapter 4, we develop a logic of action and norms (LAN) to reason about instrumentality statements. We identify a ubiquitous yet previously unaddressed norm category called *norms of instrumentality*, formalize it, and investigate its logical relations to other well-known norm categories. Based on the work of von Wright, we discuss possible extensions of LAN that model more refined instrumentality notions.

In Chapter 5, we provide an application of the logical language developed in Chapter 4 to Sanskrit philosophy. In particular, we formally analyze the deontic theory of the Mīmāṃsā philosopher Maṇḍana, which reduces all commands to statements about actions as instruments for (un)desirable results. We provide a sound and complete logic (LM) capturing this reduction and use the logic to enhance our understanding of Maṇḍana's theory. We show how the logic LM deals with well-known deontic paradoxes.

In Part III, we use methods from formal argumentation to address deontic explanations and defeasible normative reasoning.

In Chapter 6, we introduce a modular proof theoretic formalism that accommodates explanation by integrating meta-reasoning about the (in)applicability of norms into the object language of its proofs. The resulting calculi are called Deontic Argumentation Calculi (DAC). Using these calculi, we provide a sound and complete argumentative characterization of the class of nonmonotonic constrained Input/Output logics. We discuss the explanatory nature of our formalism by applying existing explanation methods from the argumentation literature to our formalism.

In Chapter 7, we develop Annotated Calculi (AC), a highly modular class of proof systems internalizing aspects of formal argumentation within the object language of its proofs. We show the consequence relation of AC to be nonmonotonic and to strongly correspond with the inference relation of formal argumentation. We extend the formalism to include defeasible normative reasoning. Namely, we incorporate the DAC formalism of Chapter 6 and show that a correspondence with formal argumentation is preserved.

Last, in Chapter 8, we conclude by giving an overview of the central contributions of this thesis. We reflect on general conclusions that can be drawn from the conducted research and discuss the most promising future research directions.

1.4 Publications

The chapters comprising Parts I–III of this dissertation are extensions of published, peer-reviewed articles. In each respective chapter, we explain in detail the differences between these articles and the present work. Here, we briefly list the relevant publications.

Chapter 2

- Kees van Berkel and Tim Lyon (2019). “A Neutral Temporal Deontic STIT Logic”. In: *Logic, Rationality, and Interaction - 7th International Workshop (LORI 2019).*

Chapter 3

- Kees van Berkel and Tim Lyon (2021). “The Varieties of Ought-Implies-Can and Deontic STIT Logic”. In: *Deontic Logic and Normative Systems - 15th International Conference (DEON 2021).*

Chapter 4

- Kees van Berkel, Tim Lyon, and Francesco Olivieri (2020). “A Decidable Multi-agent Logic for Reasoning About Actions, Instruments, and Norms”. In: *Logic and Argumentation - Third International Conference (CLAR 2020).*
- Kees van Berkel and Matteo Pascucci (2018). “Notions of instrumentality in agency logic”. In: *International Conference on Principles and Practice of Multi-Agent Systems (PRIMA 2018).*

Chapter 5

- Kees van Berkel, Agata Ciabattoni, Elisa Freschi, Francesca Gulisano, and Maya Olszewski (2021). “The Gentle Murder Paradox in Sanskrit Philosophy”. In: *Deontic Logic and Normative Systems - 15th International Conference (DEON 2021).*
- Kees van Berkel, Agata Ciabattoni, Elisa Freschi, Francesca Gulisano, and Maya Olszewski (2022) “Deontic paradoxes in Mīmāṃsā logics: there and back again”. In: *Journal of Logic, Language, and Information.*[14]

[14] This journal publication is an extended version of the preceding conference article.

Chapter 6

- Kees van Berkel and Christian Straßer (2022). “Reasoning With and About Norms in Logical Argumentation”. In: *Frontiers in Artificial Intelligence and Applications: Computational Models of Argumen (COMMA 2022).*

Chapter 7

- Ofer Arieli, Kees van Berkel, and Christian Straßer. “Annotated Sequent Calculi for Paraconsistent Reasoning and Their Relations to Logical Argumentation”. In: *Main Track of the 31st International Joint Conference on Artificial Intelligence (IJCAI 2022).*

Part I

Agency and Normative Reasoning

CHAPTER 2

Time, Choice, and Obligation

This chapter deals with obligation and choice in an explicitly temporal setting. We focus on the logic of 'Seeing To It That' (for short, STIT), a prominent formalism employing modal logics to represent and analyze agentive choice in indeterministic time. Since its beginning, STIT logic has been investigated in the light of deontic modalities, such as obligations, prohibitions, and permissions. What is more, deontic STIT scenarios have been extensively discussed against the background of temporal structures (Bartha, 1993; Belnap, 1991; Belnap et al., 2001; Horty and Belnap, 1995). Most notable is Horty's highly influential work '*Agency and Deontic Logic*' (2001). Horty argues that the temporal multi-agent setting provides good reasons for adopting a more refined notion of obligation in STIT. The resulting obligation is called the *dominance ought*. Horty's arguments concern the interaction between deontic modalities and the *implicit* underlying semantic framework of indeterministic time. Surprisingly, a logic of *explicit* temporal deontic reasoning in STIT has not yet been developed.[1] In this chapter, we set out to do this. Our intentions are twofold: we want to formally reassess some of the arguments given by Horty and further our understanding of obligations in the context of time and agency. The first objective is, thus, phrased:

Objective 1. *Develop a sound and complete Temporal Deontic* STIT *logic.*

We propose a sound and complete logic called Temporal Deontic STIT logic, referred to as TDS_n. The logic is a synthesis of various systems in the literature: non-deontic basic STIT logic (Belnap et al., 2001) extended with deontic modalities (Horty, 2001) and the temporal characterization of STIT (Lorini, 2013). The latter faithfully represents the implicitly temporal structures of the traditional STIT semantic: Branching Time frames (BT) with Agentive Choice functions, for short, BT+AC-frames. For a philosophical discussion of BT+AC frames, we refer to the work of Belnap and Perloff (1988).

[1] Broersen (2008) and Lorini (2013) provide temporal STIT logics in which obligations are *defined* using violation constants (without providing corresponding axiomatizations).

The semantics of Horty's (2001) deontic STIT logic of dominance ought is based on utility assignments. In the sequel, we call this the *Utilitarian* STIT tradition. Initially formulated by Jeremy Bentham (1789/1996), the influential theory of utilitarianism has promoted *utility calculation* as a ground for ethical deliberation: e.g., act utilitarianism classifies acts as morally right or wrong based on their comparative utility. For this reason, utility-based approaches to formal normative reasoning are promising (Åqvist, 1969). Unfortunately, each available utility function has its own (dis)advantages that give rise to philosophical puzzles—several of them addressed by Horty (2001). To avoid such problems, we provide an alternative semantic account of TDS_n by adopting relational semantics.

Objective 2. *Provide a modular characterization of Temporal Deontic* STIT *logic through relational semantics, bypassing the use of utility functions.*

An essential advantage of using relational semantics is its modularity. In our case, it facilitates a better understanding of the semantic properties of the involved deontic operators. (Furthermore, a relational characterization facilitates the formalization of a wider variety of alternative deontic properties. We demonstrate this in Chapter 3.) An immediate question arising from the above objective is whether the relational characterization of deontic STIT is equivalent to its utilitarian characterization.

Objective 3. *Formally investigate the relation between Utilitarian* STIT *semantics and Deontic* STIT *semantics.*

In this chapter, we develop a translation that enables us to constructively transform the relational STIT semantics into utilitarian STIT semantics—while preserving satisfiability—thus recovering the traditional utilitarian approach developed by Horty (2001).

Last, Horty (2001) extensively investigates various utility functions in the light of branching time frames and provides good reasons for preferring certain utility functions over others. However, the logical language used is atemporal, and Murakami (2005) proved that the corresponding atemporal deontic STIT logic cannot logically differentiate between the various types of utility assignments discussed by Horty. Furthermore, it was left as an open question to investigate "how various operators for deontic notions behave and interact in a temporal structure" (Murakami, 2005, p.5). Employing the developed TDS logic, we formally reassess some of the observations made by Horty (2001) concerning deontic STIT logic.

Objective 4. *Investigate whether the temporal extension of deontic* STIT *has consequences for the use of utility functions.*

In particular, we are interested in whether there is a formal difference between utility functions that restrict the assignment of utilities to single moments in time and those that consider the overall utility of a complete branch of a branching time structure, called a history. In this chapter, we argue that some utility functions—equivalent in a non-temporal setting—not only differ in a temporal setting but cause substantial problems with respect to contrary-to-duty (CTD) scenarios and deliberative agency.

Contributions. In this chapter, we address the above four objectives. We make four main contributions. First, we fill a long-standing gap in the STIT literature by providing a sound and weakly complete temporal deontic STIT logic. We do this by employing relational semantics. In this respect, our approach extends the results by Balbiani et al. (2008) by showing that (temporal) deontic STIT logics can likewise be characterized without using the traditional BT+AC frames.

Second, we prove several equivalence results between the two semantic approaches. We provide a constructive transformation between models adopting relational semantics and those using utility-based semantics. For instance, we observe that the language of atemporal deontic STIT is not expressive enough to differentiate between binary utility assignments and those grounded in the set of natural numbers.

Third, all of the above results hold for temporal deontic STIT logic and atemporal deontic STIT logic. For the latter, we additionally prove that the logic is strongly sound and complete.

Fourth, the increase of expressivity gained by extending deontic STIT with temporal modalities provides interesting insights into the use of utility functions. Namely, we demonstrate that certain utility functions, equivalent in an atemporal deontic STIT setting, are no longer equivalent for its temporal extension. From a philosophical point of view, we argue that two-valued utility assignments that assign utilities to complete histories are unsuitable for deliberative agency and contrary-to-duty reasoning in temporal settings. From a technical point of view, we prove that the logic TDS_n is *incomplete* for temporal STIT frames adopting these two-valued utility functions.

Differences. The results presented in this chapter were first published in (van Berkel and Lyon, 2019a). Novel contributions are the following: We provide a different axiomatization of the obligation modality $\otimes_i$ and show (Lemma 2.12) that the resulting axioms are equivalent to the axioms employed by Murakami (2005). We give the complete proofs of all the results in (van Berkel and Lyon, 2019a) and show that the results extend to the atemporal deontic STIT logic DS_n. Last, in (van Berkel and Lyon, 2019a), we informally argued that certain utility functions ranging over histories cause problems in an explicit temporal setting. Here, we make this formally precise by proving the incompleteness of TDS_n with respect to temporal utilitarian STIT-frames employing two-valued utility assignments ranging over complete histories.

Outline. In Section 2.1, we introduce the temporal deontic STIT logic TDS_n and its atemporal subsystem DS_n. Thereafter, in Section 2.2, we prove soundness of both logics and demonstrate that TDS_n is weakly complete and that DS_n is strongly complete. In Section 2.3, we prove equivalence results between the relational semantics of TDS_n (and DS_n) and the utilitarian STIT semantics. We then discuss the problem of employing two-valued utility functions in an explicitly temporal STIT setting and provide an incompleteness result in Section 2.4. Last, we relate our work to the literature and point out future work in Section 2.5.

2.1 Temporal Deontic STIT Logic

In this section, we introduce the $\underline{\text{T}}$emporal $\underline{\text{D}}$eontic $\underline{\mathsf{S}}\mathsf{TIT}$ logic, referred to as TDS_n (Objective 1). We provide a Hilbert-style axiomatization of TDS_n and a corresponding semantic characterization using relational semantics (Objective 2). Due to the modularity of our approach, we simultaneously introduce the *atemporal* $\underline{\text{D}}$eontic $\underline{\mathsf{S}}\mathsf{TIT}$ logic DS_n as a proper subsystem of the former. We start with an informal discussion of the language employed.

Indeterministic Time. Agency presupposes choice. Choice presupposes indeterministic time. This is a notion of time in which the future is open—i.e., not fully determined—and influenceable by the choices that agents make. A *moment* is then a point in time at which agents exercise choices that affect the possible continuations of time. Although the past of a given moment is uniquely determined by the course of events that led to that moment, at that moment, several futures are still possible. In other words, one may think of indeterminism as a branching time structure represented as a *tree*: the past is rooted in a linear sequence of moments, whereas the future branches out. Given such a branching time structure, each possible timeline of consecutive moments is called a *history*. In the sequel, we use timeline and history interchangeably. In other words, a moment in a branching time structure is a point in time where previously indistinguishable histories split, possibly through the influence of agents. In order to refer to the past and the future, we use the modal operators H and G, respectively. The former expresses that “it has always been that” (some proposition holds) and the latter that “it will always be that” (some proposition holds). Let P and F be the duals of H, respectively G (Prior, 1967) expressing that “somewhere in the past” (some proposition holds), respectively “somewhere in the future” (some proposition holds). We refer to the work of Belnap and Perloff (1988), Belnap et al. (2001), and Thomason (1984) for extensive discussions of indeterminist time.

Agents. Choices are exercised by agents. We denote agents by numbers $i \in \mathbb{N} = \{1, 2, 3, \dots\}$. This chapter focuses on multi-agent settings that take agents as individuals. Nature may also be considered an agent (von Wright, 1963a). We do not discuss choices made by arbitrary groups of agents (Herzig and Schwarzentruber, 2008). The only exception is the grand coalition of agents, which is used to characterize the outcome of all agents acting together (Lorini, 2013).

Choices. Different choices may be available to different agents at different moments in time. The characteristic feature of basic STIT logic is the use of an instantaneous *choice* operator $[i]$ for each agent i, which informally expresses that “agent i sees to it that” (some proposition holds). The operator is instantaneous in the sense that choice refers to what an agent can directly see to at a given moment in time.[2] In a multi-agent world, a single agent cannot uniquely determine the future by acting. For instance, when I decide to go to a concert, it may be that my friend joins me but also that she stays at home. Nevertheless, that I see to it that I go to the concert excludes a future continuation of this moment where I stayed at home. Hence, what an agent can do via exercising choice is to constrain or limit the possible courses of events. In other words, STIT models agency in indeterministic time under *uncertainty of choice*. We interpret the dual $\langle i \rangle$ of $[i]$ as "agent i sees possibly to it that" (some proposition holds). The position of "possibly" denotes that the proposition can be a consequence of the agent's choice (although this might not be guaranteed through the choice alone). Last, the modal operator $[Ag]$ represents that "the grand coalition of agents sees to it that" (some proposition holds).

Settledness. At any given moment, there are states of affairs that cannot be altered by any of the agents' (joint) choices. Such states of affairs are *settled true* at the moment in question. Basic STIT logic includes a *settledness* operator $\Box$ to refer to such states of affairs. For instance, the formula $\Box\texttt{tuesday}$ states that "at this moment, it is settled true that it is Tuesday." In basic STIT logic, this implies that no choice is available to any of the agents to see to it that today is *not* Tuesday. In such cases, we sometimes say that $\texttt{tuesday}$ is realized independently of any of the agents' choices. The dual operator $\Diamond$ expresses that some state of affairs is possible or realizable. The settledness operator plays an essential role in characterizing the relations between different choices of agents. For instance, $\Diamond[i]\texttt{concert}$ expresses that agent i has a choice to attend the concert.

Deliberative choices. The above language enables the construction of complex formulae such as $[i]\texttt{concert} \land \neg\Box\texttt{concert}$ which informally expresses that agent i sees to it that she attends the concert although it is not settled true that she will attend. In fact, this formula is an instance of the defined *deliberative* STIT operator, i.e., $[i]^d\varphi := [i]\varphi \land \neg\Box\varphi$.[3] Deliberative choices capture the idea that whenever an agent sees to it that φ, it is not necessarily the case that φ (Horty and Belnap, 1995).

Obligations. Choices lead to different continuations of time, and obligations prescribe certain choices over others. In the context of STIT, obligation is an *agentive* modality $\otimes_i$ for each agent i. Belnap and Perloff (1988) propose the canonical reading of $\otimes_i$ as "agent i is obligated to see to it that" (some proposition holds), whereas Horty (2001) interprets $\otimes_i$ as "agent i ought to see to it that" (some proposition holds). We use both interchangeably. These proposed readings include the agentive 'see to it that'. Belnap and Perloff (1988) argue that $\otimes_i$ is only *quasi-agentive* since although it involves both an agent and an agentive, the agent is tied to the normative (i.e., ought) and not to the agentive (i.e., 'see to it that'). To illustrate, the formula $\otimes_i\texttt{concert}$ is informally read

[2] The operator $[i]$ is also referred to as the *Chellas* STIT modality (Belnap et al., 2001). Alternative non-instantaneous STIT operators are the *achievement* STIT referring to the past and alternative courses of events (Belnap et al., 2001), and the *next* STIT referring to future moments as the result of agential choice (Broersen, 2011a).

[3] Xu (1998) provides a sound and complete characterization of the deliberative STIT operator taken as a primitive modality. The idea to combine two modal operators in order to define a (non-normal) deliberative choice operator was also adopted by Elgesem (1997), Kanger (1972), and Pörn (1977, Ch.1).

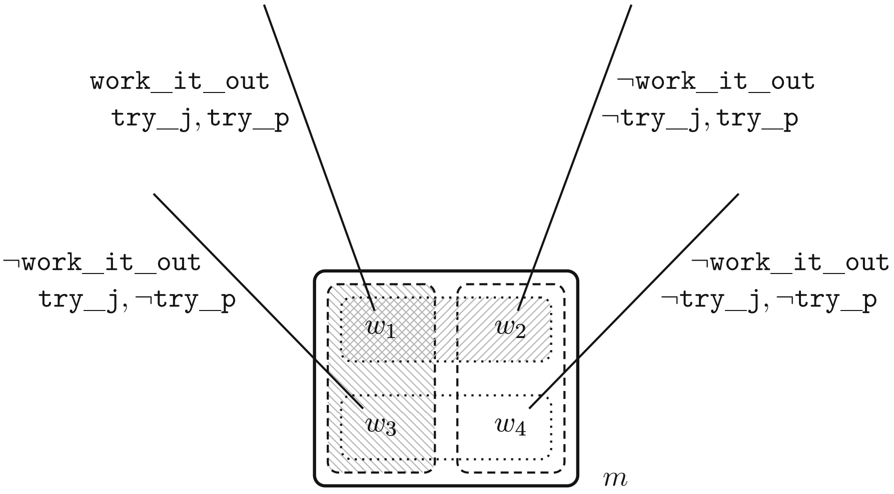


Figure 2.1: A graphical illustration of the single-moment scenario in Example 2.1.

as “agent i ought to see to it that she attends the concert” (e.g., because she made a promise). In Chapter 3, we argue that weaker interpretations of the $\otimes_i$ operator are possible and even desirable.

Example 2.1 (A single-moment two-agent example)**.** *Consider a two-agent scenario. Let John and Paul be the two agents, i.e.,* $\mathsf{Agents} = \{j, p\}$*. Suppose John and Paul got into a fight, and now each of them is faced with two choices: they can each try to work it out, or they can decide not to work it out. Let* $\mathtt{try\_j}, \mathtt{try\_p}$*, and* $\mathtt{work\_it\_out}$ *stand for “John tries to work it out”, “Paul tries to work it out”, and “it works out”. The choices are formalized as follows:*

$$\Diamond[j]\mathtt{try\_j} \textit{ and } \Diamond[j]\neg\mathtt{try\_j};$$

$$\Diamond[p]\mathtt{try\_p} \textit{ and } \Diamond[p]\neg\mathtt{try\_p}.$$

Furthermore, assume it is possible that John and Paul work it out, i.e., $\Diamond\mathtt{work\_it\_out}$*. We stipulate that this can* only *be the case if both agents try to do so, i.e.,* $\Box(\mathtt{work\_it\_out} \rightarrow ([j]\mathtt{try\_j} \wedge [p]\mathtt{try\_p}))$*. Let the moment* m *consist of the four possible combinations of the above choices. Following Balbiani et al. (2008), we interpret a moment as a collection of* worlds*, where each world corresponds to a possible continuation of time determined by the joint choices of the involved agents. Figure 2.1 gives a graphical representation of the scenario. The moment* m *consists of four possible continuations* w_1, w_2, w_3*, and* w_4*. Both agents have two choices at* m*. The two choices of* j *limit the future to either* $\{w_1, w_3\}$ *or* $\{w_2, w_4\}$ *and are graphically represented by ‘- - -’ lines. The two choices of* p *restrict the future to either* $\{w_1, w_2\}$ *or* $\{w_3, w_4\}$ *and are graphically represented by*

‘· · ·’ lines. Clearly, j and p can together see to it that they work it out by exercising the choices $\{w_1, w_3\}$ (i.e., $[j]\mathtt{try\_j}$) and $\{w_1, w_2\}$ (i.e., $[p]\mathtt{try\_p}$), leading to the unique continuation w_1 at which they work it out (i.e., $\mathtt{work\_it\_out}$). The four vertical solid lines ‘——’ denote the four distinct histories (i.e., timelines) resulting from the four possible combined choices of agents j and p. Figure 2.1 furthermore illustrates that each agent can individually and deliberately guarantee that their fight is not resolved, i.e., $\Diamond[i]^d\neg\mathtt{work\_it\_out}$ (for $i \in \{j, p\}$), but only together they can work it out, i.e., $\Diamond[Ag]\mathtt{work\_it\_out}$.

Last, suppose that, at minimum, John and Paul are under the obligation to try and work it out, i.e., their obligations are $\otimes_j\mathtt{try\_j}$ and $\otimes_p\mathtt{try\_p}$. In Figure 2.1, the two choices $\{w_1, w_3\}$ (i.e., $[j]\mathtt{try\_j}$) and $\{w_1, w_2\}$ (i.e., $[p]\mathtt{try\_p}$) are shaded to denote that they are obligatory choices. In fact, we see that only if both agents comply with their duties will they work it out. In Example 2.2, we discuss an extended example that involves temporal reasoning.

Two agency principles. To classify time as an agentive branching time structure, specific properties must be met. Traditionally, the STIT formalism contains two such principles: *independence of agents* (IoA, for short) and *no choice between undivided histories* (NCbUH, for short). Both principles are accredited to Von Kutschera (1986; 1993). The IoA principle stipulates that no agent can block another agent from exercising an available choice. For instance, if I have the choice to attend a concert tonight, I can exercise this choice irrespective of any of the choices made by the other agents. Hence, one may think of IoA as a principle that characterizes choice as choice proper (in contrast to a defeasible reading of choice). From a logical point of view, IoA ensures that any combination of choices made by the agents is consistent. The NCbUH principle stipulates that if two histories remain undivided at the next moment, no agent has a choice that realizes one history but excludes the other. This principle ensures the temporal coherence of choice. We refer to Belnap et al. (2001) for a philosophical and logical discussion of these principles. The notion of the grand coalition of agents is formally employed to characterize NCbUH (see Definition 2.2 below).[4]

We define the temporal deontic STIT language $\mathcal{L}_n^{td}$ as the combined languages of atemporal deontic STIT (Horty, 2001) and temporal non-deontic STIT (Lorini, 2013). We define $\mathcal{L}_n^{d}$ as the atemporal fragment of $\mathcal{L}_n^{td}$. We use the subscript n to denote the number of agents in the formalism.

Definition 2.1 (The Languages $\mathcal{L}_n^{td}$ and $\mathcal{L}_n^{d}$)**.** *Let* Atoms $= \{p, q, r, \dots\}$ *be a denumerable set of propositional atoms and let* Agents $= \{1, 2, \dots, n\}$ *be a finite set of agent labels. The* temporal deontic STIT *language* $\mathcal{L}_n^{td}$ *is given by the following BNF grammar:*

$$\varphi ::= p \mid \neg\varphi \mid \varphi \wedge \varphi \mid \Box\varphi \mid [i]\varphi \mid \otimes_i \varphi \mid [Ag]\varphi \mid \mathsf{G}\varphi \mid \mathsf{H}\varphi$$

[4] Additionally, one may adopt the *limited choice* principle (Belnap et al., 2001). The principle restricts each agent to a maximum number of choices at each moment. We leave it to future work to extend the logics of this chapter with the limited choice principle.

where $p \in \mathsf{Atoms}$ *and* $i \in \mathsf{Agents}$*. The* atemporal deontic STIT *language* $\mathcal{L}_n^d$ *is defined as the* $\{[Ag], \mathsf{G}, \mathsf{H}\}$*-free fragment of* $\mathcal{L}_n^{td}$*.*

In what follows, we use lowercase Roman letters $p, q, r, \ldots$ to denote propositional variables and lowercase Greek letters $\varphi, \psi, \gamma, \ldots$ to denote arbitrary formulae of $\mathcal{L}_n^{td}$ and $\mathcal{L}_n^d$. We write $\mathsf{Atoms}(\varphi)$ to denote the set of atoms occurring in a formula φ. Furthermore, we use upper case Greek letters Δ, Γ, Σ to refer to arbitrary sets of $\mathcal{L}_n^{td}$ ($\mathcal{L}_n^d$) formulae. We adopt a classical propositional base logic for the logics TDS_n and DS_n. For that reason, it suffices to take the connectives $\neg$ and $\wedge$ as primitives expressing 'not', respectively 'and'. The other logical connectives for 'disjunction', 'material implication', and 'material equivalence' are defined as usual: $\varphi \vee \psi := \neg(\neg\varphi \wedge \neg\psi)$, $\varphi \rightarrow \psi := \neg\varphi \vee \psi$, and $\varphi \equiv \psi := (\varphi \rightarrow \psi) \wedge (\psi \rightarrow \varphi)$. We define tautology and contradiction as $\top := p \vee \neg p$, respectively $\bot := \neg\top$. Last, the dual operators are defined as $\langle\alpha\rangle\varphi := \neg[\alpha]\neg\varphi$ for each pair $(\langle\alpha\rangle, [\alpha]) \in \{(\Diamond, \Box), (\langle i\rangle, [i]), (\ominus_i, \otimes_i), (\langle Ag\rangle, [Ag]), (\mathsf{F}, \mathsf{G}), (\mathsf{P}, \mathsf{H})\}$. We adopt the usual notational conventions concerning brackets.

2.1.1 Axiomatization of Temporal Deontic STIT logic

The Hilbert-style axiomatization of the temporal deontic STIT logic TDS_n is given in Definition 2.2 below. We identify the atemporal deontic STIT logic DS_n as a proper subsystem of TDS_n. The axiomatization of TDS_n combines the temporal non-deontic STIT logic from Lorini (2013) with the deontic STIT logic from Murakami (2005). The deontic axioms A10 and A13 of Definition 2.2 differ from those presented by Murakami (2005) but in Section 2.3 we prove (Lemma 2.12) that the two axiomatizations are equivalent. We refer to the work of Horty (2001) and Lorini (2013) for a more detailed discussion of the axioms. Here, we discuss each axiom briefly.

All the modalities of the language are normal modal operators by virtue of the distribution axioms A4, A1, A9, A14, A18, and A21, the rule R1 and for $[i]$, $\otimes_i$, and $[Ag]$ the axioms A7, A12, respectively A17. It can be straightforwardly checked that the latter three axiom schemes, together with R1, imply necessitation for $[i]$, $\otimes_i$, and $[Ag]$.

Concerning basic STIT, $\Box$, $[i]$, and $[Ag]$ are S5 operators by virtue of A2-A3, A5-A6, respectively A15-A16. The S5 characterization of these modalities ensures that $\Box$ refers to *moments*, and that $[i]$ and $[Ag]$ refer to *choices*. We come back to these interpretations in detail when we provide the corresponding semantics on page 33. Axiom A7 expresses that whatever is settled true at a moment is also seen to by each agent at that moment. Phrased differently, if it is settled true that φ holds at a given moment, then irrespective of the choices made by any of the agents, φ holds. Axiom A8 corresponds to the independence of agents principle, i.e., any combination of agents' choices is jointly realizable. Last, axiom A17 captures the idea that all agents acting together implies the grand coalition of agents acting.

Concerning the deontic axioms, A10 expresses the idea that obligations are settled at the level of the moment. Namely, obligations express which continuations of the present moment are ideal for that agent. For that reason, obligations that hold at a given moment do not depend on the choices made by any of the agents at that moment. Axiom A11 represents the principle of *Ought implies Can*, which ensures that what an agent is obliged to see to, the agent has the choice to see to (Chapter 3 is devoted to the analysis of Ought implies Can in the context of STIT). Axiom A12 is a bridge axiom stating that everything which is settled true is also obligatory (consequently excluding obligations to bring about states of affairs that cannot be realized). Last, A13 expresses the quasi-agentive reading of $\otimes_i$ that whenever an agent has an obligation concerning φ, the agent ought to see to it that φ holds (see Remark 2.1).

Definition 2.2 (Axiomatization of TDS_n and DS_n)**.** *We define the* Hilbert-Axiomatization of TDS_n *to be the following collection of axiom schemes and rules:*

A0. *All classical propositional tautologies;*

R0. *From φ and $\varphi \to \psi$, infer ψ;*

A1. $\Box(\varphi \to \psi) \to (\Box\varphi \to \Box\psi)$;

A2. $\Box\varphi \to \varphi$;

A3. $\Diamond\varphi \to \Box\Diamond\varphi$;

A4. $[i](\varphi \to \psi) \to ([i]\varphi \to [i]\psi)$;

A5. $[i]\varphi \to \varphi$;

A6. $\langle i\rangle\varphi \to [i]\langle i\rangle\varphi$;

A7. $\Box\varphi \to [i]\varphi$;

A8. $\bigwedge_{i\in\mathsf{Agents}} \Diamond[i]\varphi_i \to \Diamond(\bigwedge_{i\in\mathsf{Agents}}[i]\varphi_i)$;

A9. $\otimes_i(\varphi \to \psi) \to (\otimes_i\varphi \to \otimes_i\psi)$;

A10. $\otimes_i\varphi \to \Box \otimes_i \varphi$;

A11. $\otimes_i\varphi \to \Diamond[i]\varphi$;

A12. $\Box\varphi \to \otimes_i\varphi$;

A13. $\otimes_i\varphi \to \otimes_i[i]\varphi$;

R1. *From φ, infer $\Box\varphi$;*

A14. $[Ag](\varphi \to \psi) \to ([Ag]\varphi \to [Ag]\psi)$;

A15. $[Ag]\varphi \to \varphi$;

A16. $\langle Ag\rangle\varphi \to [Ag]\langle Ag\rangle\varphi$;

A17. $\bigwedge_{1\leq i\leq n}[i]\varphi_i \to [Ag]\bigwedge_{1\leq i\leq n}\varphi_i$;

A18. $\mathsf{G}(\varphi \to \psi) \to (\mathsf{G}\varphi \to \mathsf{G}\psi)$;

A19. $\mathsf{G}\varphi \to \mathsf{G}\mathsf{G}\varphi$;

A20. $\mathsf{G}\varphi \to \mathsf{F}\varphi$;

A21. $\mathsf{H}(\varphi \to \psi) \to (\mathsf{H}\varphi \to \mathsf{H}\psi)$;

A22. $\varphi \to \mathsf{G}\mathsf{P}\varphi$;

A23. $\varphi \to \mathsf{H}\mathsf{F}\varphi$;

A24. $\mathsf{F}\mathsf{P}\varphi \to \mathsf{P}\varphi \vee \varphi \vee \mathsf{F}\varphi$;

A25. $\mathsf{P}\mathsf{F}\varphi \to \mathsf{P}\varphi \vee \varphi \vee \mathsf{F}\varphi$;

A26. $\mathsf{F}\Diamond\varphi \to \langle Ag\rangle\mathsf{F}\varphi$;

R2. *From φ, infer $\mathsf{G}\varphi$ and $\mathsf{H}\varphi$;*

R3. *From $(\Box\neg p \wedge \Box(\mathsf{G}p \wedge \mathsf{H}p)) \to \varphi$ with $p \notin \mathsf{Atoms}(\varphi)$, infer φ;*

where we have a copy of A4–A13 *for each $i \in \mathsf{Agents}$. The* logic TDS_n *is the smallest set of formulae from $\mathcal{L}_n^{td}$ closed under all instances of the axiom schemes and applications of the inference rules* R0–R3*. Whenever $\varphi \in \mathsf{TDS}_n$, we say that $\varphi \in \mathcal{L}_n^{td}$ is a* TDS_n-theorem *and write $\vdash_{\mathsf{TDS}_n} \varphi$.*

We define the Hilbert-Axiomatization of DS_n *to consist of the axiom schemes and rules of the left column (above), i.e., axiom schemes* A0–A13 *(for each* $i \in \mathsf{Agents}$*) and the rules* R0–R1*. The* logic DS_n *and* DS_n-theoremhood *are defined as above.*

Concerning the axiomatization of time, A18-A20 capture the common conception of branching time as transitive and serial, i.e., G is a KD4 modality. Axioms A22 and A23 serve as the central axioms of minimal temporal logic and ensure that the past (i.e., H) is the converse of the future (i.e., G), e.g., see the work of Thomason (1984). For instance, A22 expresses that what is the case now, will always going to be somewhere in the past. Furthermore, since we are dealing with branching time structures containing histories, the axioms A24 and A25 capture the idea that histories are *linear* timelines. Axiom A26 characterizes the no choice between undivided histories principle. In fact, the main reason why the grand coalition operator $[Ag]$ is added to the language $\mathcal{L}_n^{td}$ is because it enables the axiomatization of this pivotal STIT principle.

Last, the rule R3 is a variation of the *irreflexivity rule* proposed by Gabbay et al. (1994). The rule ensures that moments in a branching time structure are irreflexive and, consequently, so is time. The rule R3 is not immediately intuitive. To see how it works, consider a simplification of the rule as proposed by Gabbay et al. (1994):

$$\mathsf{R3}^* \text{ From } (\neg p \wedge [\alpha]p) \rightarrow \varphi \text{ with } p \notin \mathsf{Atoms}(\varphi), \text{ infer } \varphi.$$

where $[\alpha]$ is an arbitrary normal modal operator. This rule ensures that $[\alpha]$ behaves as an irreflexive modality. Adding the reflexivity axiom $[\alpha]\varphi \rightarrow \varphi$ to any consistent logic containing R3* would render the logic inconsistent. Namely, from $[\alpha]p \rightarrow p$ we straightforwardly obtain $([\alpha]p \wedge \neg p) \rightarrow \bot$ and, thus, by R3* we have $\bot$. At a minimum, the presence of the rule tells us that the logic does not permit any reflexive behavior. How R3 actually ensures irreflexivity of TDS_n-frames is best understood by considering the proofs of soundness (Theorems 2.1) and completeness (Theorem 2.3) in Section 2.2. We refer to Gabbay et al. (1994) for a more general discussion of the irreflexivity rule.

Definition 2.3 (TDS_n and DS_n derivations)**.** *Let* $\varphi \in \mathcal{L}_n^{td}$ *and* $\Gamma \subseteq \mathcal{L}_n^{td}$*, we define a* derivation φ *from premises* Γ *in* TDS_n*, written* $\Gamma \vdash_{\mathsf{TDS}_n} \varphi$*, as follows: there exists a sequence* $\varphi_1, \ldots, \varphi_n \in \mathcal{L}_n^{td}$ *of formulae such that* $\varphi_n = \varphi$ *and for each* $1 \leq i \leq n$*,* φ_i *is either a* TDS_n*-theorem, an assumption from* Γ*, or a consequence of an application of* R0 *to some* $\varphi_j = \psi$ *and* $\varphi_k = \psi \rightarrow \varphi_i$ *with* $j, k < i$*. A derivation in* DS_n *is defined similarly.*

Remark 2.1 (Quasi-Agentive Obligation)**.** *We point out that the logic* DS_n *defines the* quasi-agentive *reading of* $\otimes_i$*, i.e., "agent i ought to see to it that" (some proposition holds). The* DS_n*-theorem* $\otimes_i\varphi \equiv \otimes_i[i]\varphi$ *expresses this. Clearly, the left-to-right direction follows directly from* A13*. The right-to-left direction is proven as follows: First, observe that* $\otimes_i([i]\varphi \rightarrow \varphi)$ *is a theorem by an application of* R1 *to* A5 *and basic modal reasoning with* A12*. Since* $\otimes_i$ *is a normal modal operator, we can infer* $\otimes_i[i]\varphi \rightarrow \otimes_i\varphi$*. In other* *words, the modal operator $\otimes_i$ receives its quasi-agentive reading from the adopted* DS_n *axiomatization. The above shows that $\otimes_i$ is only quasi-agentive if one adopts the axioms* A9, A12, *and* A13 *(on top of the basic* STIT *logic). The class of deontic* STIT *logics introduced in Chapter 3 deliberately does not imply the quasi-agentive reading of $\otimes_i$. The reason is that certain prominent alternative readings of Ought-implies-Can cannot be axiomatized with quasi-agentive deontic modality.*

2.1.2 Semantics for Temporal Deontic STIT Logic

We forgo the traditional BT+AC semantics in characterizing TDS_n. Instead, we adopt *relational semantics* (Blackburn et al., 2004). As observed by Balbiani et al. (2008), atemporal STIT logic can be semantically characterized using relational frames that model moments as sets of worlds partitioned into equivalence classes, the latter representing the choices available to the agents at the respective moments. We adopt this approach in defining TDS_n- and DS_n-frames. The semantic characterization of the temporal properties was initially proposed by (Lorini, 2013).

Definition 2.4 (Frames and Models for TDS_n and DS_n)**.** *A* Temporal Deontic STIT-frame *(for short,* TDS_n-frame*) is defined as a tuple $\mathfrak{F} = \langle W, \mathcal{R}_\Box, \{\mathcal{R}_{[i]} \mid i \in \mathsf{Agents}\}, \{\mathcal{R}_{\otimes_i} \mid i \in \mathsf{Agents}\}, \mathcal{R}_{[Ag]}, \mathcal{R}_\mathsf{G}, \mathcal{R}_\mathsf{H}\rangle$. Let $\mathcal{R}_{[\alpha]} \subseteq W \times W$ and $\mathcal{R}_{[\alpha]}(w) := \{v \in W \mid (w,v) \in \mathcal{R}_{[\alpha]}\}$ for $[\alpha] \in \mathsf{Boxes} := \{\Box, \mathsf{G}, \mathsf{H}, [Ag]\} \cup \{[i] \mid i \in \mathsf{Agents}\} \cup \{\otimes_i \mid i \in \mathsf{Agents}\}$. Let W be a non-empty set of worlds $w, v, u, \ldots$. The following holds:*

C1 *$\mathcal{R}_\Box$ is an equivalence relation*[5]*;*

C2 *For all $i \in \mathsf{Agents}$, $\mathcal{R}_{[i]}$ is an equivalence relation;*

C3 *For all $i \in \mathsf{Agents}$, $\mathcal{R}_{[i]} \subseteq \mathcal{R}_\Box$;*

C4 *For all $w \in W$ and all $u_1, \ldots, u_n \in \mathcal{R}_\Box(w)$, $\bigcap_{i \in \mathsf{Agents}} \mathcal{R}_{[i]}(u_i) \neq \emptyset$;*

C5 *$\mathcal{R}_{[Ag]}$ is an equivalence relation;*

C6 *For all $w \in W$, $\mathcal{R}_{[Ag]}(w) \subseteq \bigcap_{i \in \mathsf{Agents}} \mathcal{R}_{[i]}(w)$;*

D1 *For all $i \in \mathsf{Agents}$ and for all $w, v, u \in W$, if $v \in \mathcal{R}_{\otimes_i}(w)$ and $u \in \mathcal{R}_\Box(w)$, then $v \in \mathcal{R}_{\otimes_i}(u)$;*

D2 *For all $i \in \mathsf{Agents}$, and all $w \in W$, there exists $v \in W$ such that for all $u \in \mathcal{R}_{[i]}(v)$, $u \in \mathcal{R}_{\otimes_i}(w)$;*

D3 *For all $i \in \mathsf{Agents}$, $\mathcal{R}_{\otimes_i} \subseteq \mathcal{R}_\Box$;*

D4 *For all $i \in \mathsf{Agents}$, and all $w, v, u \in W$, if $v \in \mathcal{R}_{\otimes_i}(w)$ and $u \in \mathcal{R}_{[i]}(v)$, then $u \in \mathcal{R}_{\otimes_i}(w)$;*

[5] That is, $\mathcal{R}_\Box$ is reflexive and euclidean.

T1 *$\mathcal{R}_{\mathsf{G}}$ is a transitive and serial relation;*

T2 *$\mathcal{R}_{\mathsf{H}}$ is the converse of $\mathcal{R}_{\mathsf{G}}$, i.e., $\mathcal{R}_{\mathsf{H}} = \{(w, v) \mid (v, w) \in \mathcal{R}_{\mathsf{G}}\}$;*

T3 *For all $w, u, v \in W$, if $u \in \mathcal{R}_{\mathsf{H}}(w)$ and $v \in \mathcal{R}_{\mathsf{H}}(w)$, then $v \in \mathcal{R}_{\mathsf{H}}(u)$, $u = v$, or $u \in \mathcal{R}_{\mathsf{H}}(v)$;*

T4 *For all $w, u, v \in W$, if $u \in \mathcal{R}_{\mathsf{G}}(w)$ and $v \in \mathcal{R}_{\mathsf{G}}(w)$, then $v \in \mathcal{R}_{\mathsf{G}}(u)$, $u = v$, or $u \in \mathcal{R}_{\mathsf{G}}(v)$;*

T5 *$\mathcal{R}_{\mathsf{G}} \circ \mathcal{R}_{\Box} \subseteq \mathcal{R}_{[Ag]} \circ \mathcal{R}_{\mathsf{G}}$, where $\mathcal{R}_{[\alpha]} \circ \mathcal{R}_{[\beta]} := \{(w, v) \mid$ there is $u \in W$, $u \in \mathcal{R}_{[\alpha]}(w)$, and $v \in \mathcal{R}_{[\beta]}\}$ for $[\alpha], [\beta] \in \mathsf{Boxes}$;*

T6 *For all $w, u \in W$, if $u \in \mathcal{R}_{\Box}(w)$, then $u \notin \mathcal{R}_{\mathsf{G}}(w)$.*

A TDS_n-model *is a tuple $\mathfrak{M} = \langle \mathfrak{F}, V \rangle$ where $\mathfrak{F}$ is a TDS_n-frame and V is a valuation function mapping propositional variables to subsets of W, i.e., $V\colon \mathsf{Atoms} \mapsto \wp(W)$.*

An (atemporal) Deontic STIT-frame *(for short,* DS_n-frame*) is defined to be a tuple $\mathfrak{F} = \langle W, \mathcal{R}_{\Box}, \{\mathcal{R}_{[i]} \mid i \in \mathsf{Agents}\}, \{\mathcal{R}_{\otimes_i} \mid i \in \mathsf{Agents}\} \rangle$. Where $\mathfrak{F}$ satisfies **C1-C4** and **D1-D4**. A* DS_n-model *is a tuple $\mathfrak{M} = \langle \mathfrak{F}, V \rangle$ where $\mathfrak{F}$ is a DS_n-frame and V is a valuation function as defined above.*

In Definition 2.4, we write **Ci** ($i \in \{1, \ldots, 6\}$), **Di** ($i \in \{1, \ldots, 4\}$), and **Ti** ($i \in \{1, \ldots, 6\}$ to denote the choice properties, deontic properties, respectively temporal properties of TDS_n-frames. We discuss each property in turn.

First, observe that the relation $\mathcal{R}_{[\alpha]}$ for $[\alpha] \in \{\Box\} \cup \{[i] \mid i \in \mathsf{Agents}\} \cup \{[Ag]\}$ is an equivalence relation by **C1**, **C2**, and **C5**, and thus the set $\mathcal{R}_{[\alpha]}(w) = \{v \mid (w, v) \in \mathcal{R}_{[\alpha]}\}$ is an *equivalence class* (cf. the $\mathsf{S5}$ axiomatization of $\Box$, $[i]$, and $[Ag]$ in Definition 2.2). Property **C1** stipulates that TDS_n-frames are partitioned into $\mathcal{R}_{\Box}$-equivalence classes representing *moments*. For each agent in the language, **C2** and **C3** partition moments into equivalence classes representing the agent's *choices* at these moments (cf. $\mathsf{A7}$). In what follows, we often call $\mathcal{R}_{\Box}(w)$ a *moment* and for each $v \in \mathcal{R}_{\Box}(w)$, we refer to $\mathcal{R}_{[i]}(v)$ as a *choice* for agent i at moment $\mathcal{R}_{\Box}(w)$. Property **C4** captures the IoA principle, ensuring that the choices of agents acting simultaneously are jointly consistent (cf. $\mathsf{A8}$). Furthermore, **C5** expresses that the set $\mathcal{R}_{[Ag]}(w)$ is an equivalence class, i.e., a choice of the grand coalition of agents acting together. Last, **C6** ensures that all agents acting together is a necessary condition for the grand coalition of agents acting (cf. $\mathsf{A17}$).[6]

Deontic property **D1** ensures that obligations refer to what is obligatory at a given moment irrespective of the choices made by the agents at that moment (cf. $\mathsf{A10}$). Notice

[6] As shown by Lorini (2013), condition **C6** can be strengthened to equality: i.e., **C6*** for all $w \in W$, $\mathcal{R}_{[Ag]}(w) = \bigcap_{i \in \mathsf{Agents}} \mathcal{R}_{[i]}(w)$. In such a setting, completeness is proven by demonstrating that each TDS_n-frame can be transformed into a frame satisfying the same formulae with the strengthened condition **C6***. Hence, the language $\mathcal{L}_n^{td}$ is not expressive enough to distinguish between the two frame classes.

that obligations may still differ from moment to moment in a branching time setting. Property **D2** semantically captures the principle of Ought implies Can (cf. A11). **D3** enforces that ideal worlds are confined to moments (cf. A12). This condition implies that every ideal world is realizable at its corresponding moment. Subsequently, **D4** expresses that agent-dependent obligations are about choices, thus enforcing that every ideal world extends to a complete ideal choice (cf. A13). Property **D4** is central for the *quasi-agentive* reading of the obligation $\otimes_i$ (cf. Remark 2.1). In what follows, we sometimes refer to $\mathcal{R}_{[i]}(v) \subseteq \mathcal{R}_{\otimes_i}(w) \subseteq \mathcal{R}_{\Box}(w)$ as a deontically optimal choice for agent i at moment $\mathcal{R}_{\Box}(w)$.

Combined, the conditions **T1**–**T6** ensure that TDS_n-frames are irreflexive, temporal orderings of moments in a branching time structure. First, **T1** and **T2** ensure that for each history, the future is transitive and serial, and the past is the converse of the future. Properties **T3** and **T4** stipulate that future and past sequences of worlds are linearly ordered. As discussed a the beginning of this section, we call such a (maximally) linearly ordered sequence a *history*, representing a possible timeline in a branching time structure. Formally, we can express the history of which a world $w \in W$ is a member as the set $\mathcal{R}_{\mathsf{G}}(w) \cup \mathcal{R}_{\mathsf{H}}(w) \cup \{w\}$. Just like $\mathcal{R}_{\Box}(w)$ and $\mathcal{R}_{[i]}(w)$ refer to moments, respectively choices, we use $\mathcal{R}_{\mathsf{G}}(w)$ and $\mathcal{R}_{\mathsf{H}}(w)$ to refer to the future, respectively past history of w. Property **T6** ensures the temporal irreflexivity of moments.

In particular, condition **T5** ensures the STIT principle of no choice between undivided histories. Namely, if two histories remain undivided at the next moment, no agent has a choice that realizes one history but excludes the other. To see how **T5** formally captures this idea, suppose towards a contradiction that agent i has two choices $\mathcal{R}_{[i]}(v)$ and $\mathcal{R}_{[i]}(u)$ at a moment $\mathcal{R}_{\Box}(w)$ such that the histories of these choices are undivided at a next moment. Then, there are $v' \in \mathcal{R}_{\mathsf{G}}(v)$ and $u' \in \mathcal{R}_{\mathsf{G}}(u)$ such that $\mathcal{R}_{\Box}(v') = \mathcal{R}_{\Box}(u')$, i.e., the two future worlds v' and u' are part of the same future moment. In other words, $(v, u'), (u, v') \in \mathcal{R}_{\mathsf{G}} \circ \mathcal{R}_{\Box}$. Hence, by **T5** $(v, u'), (u, v') \in \mathcal{R}_{[Ag]} \circ \mathcal{R}_{\mathsf{G}}$. This means that there is a $z \in \mathcal{R}_{\Box}(w)$ such that $(v, z) \in \mathcal{R}_{[Ag]}$ and $(z, u') \in \mathcal{R}_{\mathsf{G}}$. By the linearity of histories, we know that $z = u$ and so $(v, u) \in \mathcal{R}_{[Ag]}$. However, by the fact that $\mathcal{R}_{[i]}(v) \cap \mathcal{R}_{[i]}(u) = \emptyset$ we know that v and u cannot be part of the same choice of the grand coalition of agents acting at $\mathcal{R}_{\Box}(w)$, i.e., $(v, u) \notin \mathcal{R}_{[Ag]}$. Contradiction. Consequently, **T5** ensures that the ordering of moments is linearly closed with respect to the past and allows for branching with respect to the future. We refer to Belnap et al. (2001) for a philosophical discussion of this principle.

The semantic interpretation of $\mathcal{L}_n^{td}$ is defined as usual.

Definition 2.5 (Semantics of TDS_n- and DS_n-models)**.** *Let $\mathfrak{M}$ be a TDS_n-model and let $w \in W$ of $\mathfrak{M}$. The* satisfaction *of a formula $\varphi \in \mathcal{L}_n^{td}$ in $\mathfrak{M}$ at w is defined accordingly:*

1. $\mathfrak{M}, w \models p$ iff $w \in V(p)$;
2. $\mathfrak{M}, w \models \neg\varphi$ iff *not* $\mathfrak{M}, w \models \varphi$;
3. $\mathfrak{M}, w \models \varphi \wedge \psi$ iff $\mathfrak{M}, w \models \varphi$ *and* $\mathfrak{M}, w \models \psi$;

4. $\mathfrak{M}, w \models \Box\varphi$ iff *for all* $u \in \mathcal{R}_{\Box}(w)$, $\mathfrak{M}, u \models \varphi$;

5. $\mathfrak{M}, w \models [i]\varphi$ iff *for all* $u \in \mathcal{R}_{[i]}(w)$, $\mathfrak{M}, u \models \varphi$;

6. $\mathfrak{M}, w \models \otimes_i\varphi$ iff *for all* $u \in \mathcal{R}_{\otimes_i}(w)$, $\mathfrak{M}, u \models \varphi$;

7. $\mathfrak{M}, w \models [Ag]\varphi$ *iff for all* $u \in \mathcal{R}_{[Ag]}(w), \mathfrak{M}, u \models \varphi$;

8. $\mathfrak{M}, w \models \mathsf{G}\varphi$ *iff for all* $u \in \mathcal{R}_{\mathsf{G}}(w)$, $\mathfrak{M}, u \models \varphi$;

9. $\mathfrak{M}, w \models \mathsf{H}\varphi$ *iff for all* $u \in \mathcal{R}_{\mathsf{H}}(w)$, $\mathfrak{M}, u \models \varphi$.

We define $\|\varphi\|_{\mathfrak{M}} = \{w \in W \mid \mathfrak{M}, w \models \varphi\}$ *as the* truth-set *of* φ *(we often omit the subscript* $\mathfrak{M}$*). We write* $\mathfrak{M}, w \not\models \varphi$ *to indicate that* not $\mathfrak{M}, w \models \varphi$.

A formula φ *is* globally true *on a* TDS_n*-model* $\mathfrak{M}$*, written* $\mathfrak{M} \models \varphi$*, if and only if* φ *is satisfied at every world* $w \in W$ *of* $\mathfrak{M}$*. A formula* φ *is* TDS_n-valid*, written* $\models_{\mathsf{TDS}_n} \varphi$*, if and only if it is globally true on every* TDS_n*-model. Last, we say that* $\Gamma \subseteq \mathcal{L}_n^{td}$ *semantically entails* φ*, written* $\Gamma \models_{\mathsf{TDS}_n} \varphi$*, if and only if for all* TDS_n*-models* $\mathfrak{M}$ *and worlds* $w \in W$ *of* $\mathfrak{M}$*, if* $\mathfrak{M}, w \models \psi$ *for all* $\psi \in \Gamma$*, then* $\mathfrak{M}, w \models \varphi$*. The logic induced by the class of all* TDS_n*-models is the set of* TDS_n*-valid formulae.*

Satisfaction of a formula $\varphi \in \mathcal{L}_n^d$ *in a* DS_n*-model is defined by clauses (1)-(6). Global truth, validity, and semantic entailment for* DS_n*-models are defined as above.*

Example 2.2 (A Temporal Deontic Scenario)**.** *To illustrate temporal deontic* STIT *models, consider an extension of the scenario in Example 2.1. Recall that the two agents John and Paul (i.e.,* $\mathsf{Agents} = \{j, p\}$*) were in a feud and are both under the obligation to try to work it out, i.e., (a)* $\otimes_j\texttt{try\_j} \land \otimes_p\texttt{try\_p}$*. Furthermore, they work it out together only if they try, i.e., (b)* $\Box(\texttt{work\_it\_out} \rightarrow ([j]\texttt{try\_j} \land [p]\texttt{try\_p}))$*. If both agents fulfill their duty and work it out, then they ought to thank each other (out of politeness), i.e., (c)* $\Box(\texttt{work\_it\_out} \rightarrow \mathcal{F}(\otimes_j\texttt{thank\_j} \land \otimes_p\texttt{thank\_p}))$*, where* thank_j *and* thank_p *express "John thanks Paul", respectively "Paul thanks John". However, if they do not manage to work it—i.e., at least one of them violating the initial obligation—they both ought to get a little help from their friends (say, for mediation), i.e., (d)* $\Box(\neg\texttt{work\_it\_out} \rightarrow \mathsf{F}(\otimes_j\texttt{help\_j} \land \otimes_p\texttt{help\_p}))$*, where* help_j *and* help_p *express "John gets a little help from his friends", respectively "Paul gets a little help from his friends". This second situation represents a temporal* contrary-to-duty *(CTD) scenario in which obligations arise from the violation of a previous obligation.*[7]

Figure 2.2 graphically represents the above scenario in a branching time TDS_n*-model. We briefly explain its representation. The model consists of a root moment* $\mathcal{R}_{\Box}(\omega^{\alpha})$ *and four immediate successor moments. To illustrate,* $\mathcal{R}_{\Box}(v_i)$ $(1 \leq i \leq 4)$ *is the moment continuing from* ω^{v}*. We stress that since, in total, four distinct histories emerge from*

[7] In Section 2.4, we discuss CTD in the context of TDS_n at length (see Chapter 1 for an introduction).

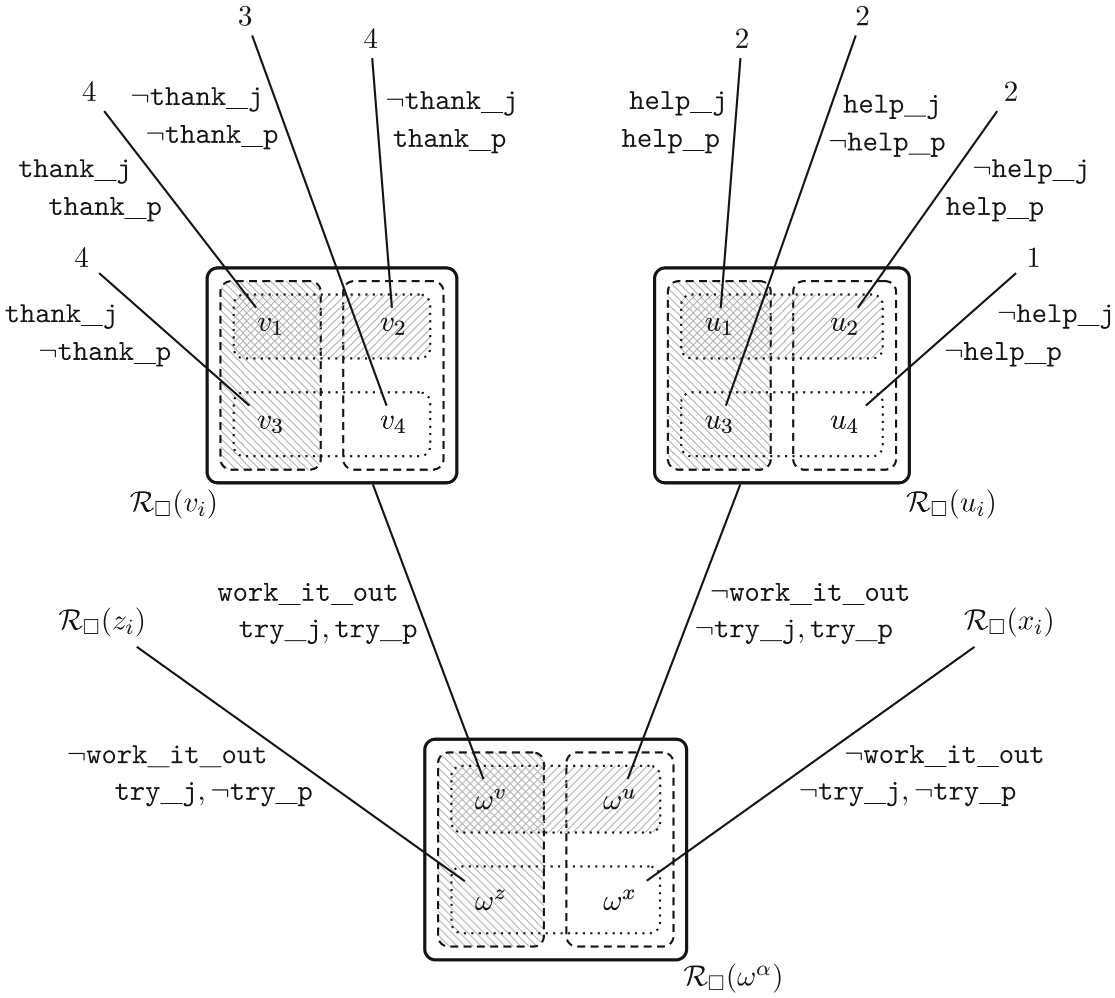


Figure 2.2: A graphical illustration of the temporal contrary-to-duty scenario in Example 2.2. For the moment $\mathcal{R}_\Box(\omega^\alpha)$ the symbol ω^α represents a set of worlds for each $\alpha \in \{v, u, z, x\}$ because each ω^α leads to a future moment with four worlds. We stipulate that after each β_i with $\beta \in \{v, u, z, x\}$ and $i \in \{1, 2, 3, 4\}$ the histories indefinitely continue with single-world moments only. Moments $\mathcal{R}_\Box(z_i)$ and $\mathcal{R}_\Box(x_i)$ have a characterization identical to that of $\mathcal{R}_\Box(u_i)$ and are, for that reason, omitted from the figure. The numbers assigned to the histories represent utilities and are discussed in Section 2.3.

$\mathcal{R}_\Box(v_i)$ we know by the linearity of timelines that ω^v represents the set of four worlds $\{w_1^v, w_2^v, w_3^v, w_4^v\}$. We write ω^v to enhance readability of Figure 2.2. The same holds for ω^u, ω^x, and ω^z. Consequently, this model consists of exactly 16 histories. In fact, due to the irreflexivity and seriality of TDS_n*-models, these histories are infinite. It suffices to stipulate that each history in the model is infinite (we provide an exact model of this example in Section 2.4). Furthermore, we assume that the moments $\mathcal{R}_\Box(z_i)$ and $\mathcal{R}_\Box(x_i)$ $(1 \leq i \leq 4)$ represent exactly the same scenario as $\mathcal{R}_\Box(u_i)$ and are, for that reason, omitted from the figure. The choices of j and p are graphically represented by ‘- - -’ lines, respectively ‘· · ·’ lines. The obligatory choices for both agents are shaded, and darker shaded when overlapping. (The utilities assigned to the histories in Figure 2.2 are explained when we discuss Utilitarian* STIT *logic in Section 2.3.)*

The formulae (a), (b), (c), and (d) hold at moment $\mathcal{R}_\Box(\omega^\alpha)$. Furthermore, the obligations to thank each other—i.e., (e) $\otimes_j$`thank_j` $\wedge \otimes_p$`thank_p`*—result from the agents' joint compliance with (a) at $\mathcal{R}_\Box(\omega^\alpha)$. Namely, (e) holds at moment $\mathcal{R}_\Box(v_i)$ which is a continuation of ω^v resulting from the joint choices $[j]$*`try_j` *and $[p]$*`try_p` *at $\mathcal{R}_\Box(\omega^\alpha)$, i.e., $\{\omega^v, \omega^u\} \cap \{\omega^v, \omega^z\}$. Similarly, John and Paul's obligations to get some help from their friends—i.e., (f) $\otimes_j$*`help_j` $\wedge \otimes_p$`help_p`*—result from either of the two agents violating their obligation at $\mathcal{R}_\Box(\omega^\alpha)$. For instance, (f) holds at the moment $\mathcal{R}_\Box(u_i)$, which is a continuation of world ω^u resulting from the joint choices $[j]\neg$*`try_j` *and $[p]$*`try_p` *at $\mathcal{R}_\Box(\omega^\alpha)$, i.e., $\{\omega^u, \omega^x\} \cap \{\omega^v, \omega^u\}$. The same reasoning applies to (f) and moments $\mathcal{R}_\Box(z_i)$ and $\mathcal{R}_\Box(x_i)$.*

2.2 Soundness and Completeness

Soundness of the logic TDS_n is obtained by demonstrating that all TDS_n axioms are TDS_n-valid and the logical rules of TDS_n preserve truth on any TDS_n-frame. This is a standard strategy for normal modal logics (Blackburn et al., 2004). In the sequel, we make (often implicit) use of the following useful lemma.

Lemma 2.1. *The following holds for any* TDS_n*- and* DS_n*-frame. Let $w, v \in W$ and $i \in$* Agents*:*

1. *For all $v \in \mathcal{R}_\Box(w)$, we have $\mathcal{R}_\Box(w) = \mathcal{R}_\Box(v)$;*
2. *For all $v \in \mathcal{R}_{[i]}(w)$, we have $\mathcal{R}_{[i]}(w) = \mathcal{R}_{[i]}(v)$;*
3. *$\mathcal{R}_\Box(w) \neq \emptyset$ and $\mathcal{R}_{[i]}(w) \neq \emptyset$;*
4. *For all $v \in \mathcal{R}_\Box(w)$, we have $\mathcal{R}_{\otimes_i}(v) = \mathcal{R}_{\otimes_i}(w)$;*
5. *$\overline{\|\varphi\|} = \|\neg\varphi\|$ and $\|\varphi\| \cap \|\psi\| = \|\varphi \wedge \psi\|$.*

Proof. Claims (1)–(3) follow from the fact that $\mathcal{R}_\Box$ and $\mathcal{R}_{[i]}$ are equivalence classes, and statement (4) follows from property **D1** of Definition 2.4. The properties of truth sets in (5) follow by basic semantic reasoning. QED

Theorem 2.1 (Soundness of TDS_n)**.** *Let* TDS_n *be the logic from Definition 2.2. For any formula* $\varphi \in \mathcal{L}_n^{td}$*, and any* $\Gamma \subseteq \mathcal{L}_n^{td}$*: if* $\Gamma \vdash_{\mathsf{TDS}_n} \varphi$*, then* $\Gamma \models_{\mathsf{TDS}_n} \varphi$*.*

Proof. First, we demonstrate the following claim:

$$(\dagger) \text{ if } \vdash_{\mathsf{TDS}_n} \varphi, \text{ then } \models_{\mathsf{TDS}_n} \varphi.$$

We prove (†) by demonstrating that all axioms are TDS_n-valid and the logical rules of TDS_n preserve truth on the respective frame class. Take an arbitrary LM-model $\mathfrak{M}$ and an arbitrary $w \in W$ of $\mathfrak{M}$. The axiom schemes A0, A1, A4, A9, A21, A18, and A14, and rules R0, R1, and R2 are valid, respectively preserve validity on any relational frame (Blackburn et al., 2004). We omit their proofs. Validity of the remaining axioms and rule R3 is shown below.

A2 Assume $\mathfrak{M}, w \models \Box\varphi$. Hence, for all $v \in \mathcal{R}_\Box(w)$ $\mathfrak{M}, v \models \varphi$ by **C1** we know that $\mathcal{R}_\Box$ is reflexive and thus $w \in \mathcal{R}_\Box(w)$. Consequently, $\mathfrak{M}, w \models \varphi$.

A3 Assume $\mathfrak{M}, w \models \Diamond\varphi$. Hence, there is a world $v \in \mathcal{R}_\Box(w)$ such that $\mathfrak{M}, v \models \varphi$. By Lemma 2.1-(i) we know that $\mathcal{R}_\Box(w) = \mathcal{R}_\Box(v)$. Therefore, we know that for all $u \in \mathcal{R}_\Box(w)$, $v \in \mathcal{R}_\Box(u)$ with $\mathfrak{M}, v \models \varphi$. Hence, by the semantic definition of $\Box$ we know that for all $u \in \mathcal{R}_\Box(w)$, $\mathfrak{M}, u \models \Diamond\varphi$ and so $\mathfrak{M}, w \models \Box\Diamond\varphi$.

A5 Similar to A2.

A6 Similar to A3.

A7 Assume $\mathfrak{M}, w \models \Box\varphi$. Hence, by the semantic definition of $\Box$ we know that for each $v \in \mathcal{R}_\Box(w)$, $\mathfrak{M}, v \models \varphi$. That is, $\mathcal{R}_\Box(w) \subseteq \|\varphi\|$. By property **C3**, $\mathcal{R}_{[i]}(w) \subseteq \mathcal{R}_\Box(w) \subseteq \|\varphi\|$ and so $\mathfrak{M}, w \models [i]\varphi$.

A8 Assume $\mathfrak{M}, w \models \bigwedge_{i\in\mathsf{Agents}} \Diamond[i]\varphi_i$, i.e., $\mathfrak{M}, w \models \Diamond[1]\varphi_1 \wedge \ldots \wedge \Diamond[n]\varphi_n$. By the semantic definition of $\Diamond$ there are $v_1, \ldots, v_n \in \mathcal{R}_\Box(w)$ such that $\mathfrak{M}, v_1 \models [1]\varphi_1$, … , $\mathfrak{M}, v_n \models [n]\varphi_n$. Therefore, for all v_i with $1 \leq i \leq n$, $\mathcal{R}_{[i]}(v_i) \subseteq \|\varphi_i\|$. By condition **C4**, we know that there is a $u \in \bigcap_{i\in\mathsf{Agents}} \mathcal{R}_{[i]}(v_i)$. Since $u \in \mathcal{R}_{[i]}(v_i)$ for each $1 \leq i \leq n$ we have $\mathfrak{M}, u \models [1]\varphi_1 \wedge \ldots \wedge [n]\varphi_n$. Consequently, by property **C3** we know that $u \in \mathcal{R}_\Box(w)$ and thus $\mathfrak{M}, w \models \Diamond \bigwedge_{i\in\mathsf{Agents}}[i]\varphi_i$.

A10 Assume $\mathfrak{M}, w \models \otimes_i\varphi$. Suppose towards a contradiction that $\mathfrak{M}, w \not\models \Box\otimes_i\varphi$. Hence, $\mathfrak{M}, w \models \Diamond\ominus_i\neg\varphi$ and so there is a $v \in \mathcal{R}_\Box(w)$ such that $\mathfrak{M}, v \models \ominus_i\neg\varphi$. Consequently, by semantic definition of $\ominus_i$, we know there is a $u \in \mathcal{R}_{\otimes_i}(v)$ with $\mathfrak{M}, u \models \neg\varphi$. By **D1** we know that $u \in \mathcal{R}_{\otimes_i}(w)$ and so, by our initial assumption, $\mathfrak{M}, u \models \varphi$ too. Contradiction.

A11 Assume $\mathfrak{M}, w \models \otimes_i\varphi$. By **D2** we know there is a $v \in \mathcal{R}_\Box(w)$ and for all $u \in \mathcal{R}_{[i]}(v)$, $u \in \mathcal{R}_{\otimes_i}(w)$. Suppose towards a contradiction that $\mathfrak{M}, w \not\models \Diamond[i]\varphi$. Consequently, $\mathfrak{M}, w \models \Box\langle i\rangle\neg\varphi$ and so $\mathfrak{M}, v \models \langle i\rangle\neg\varphi$. By semantic definition of $\langle i\rangle$ there is a $u \in \mathcal{R}_{[i]}(v)$ such that $\mathfrak{M}, u \models \neg\varphi$. However, since $u \in \mathcal{R}_{\otimes_i}(w)$ we also have $\mathfrak{M}, u \models \varphi$. Contradiction.

A12 Similar to A7.

A13 Assume $\mathfrak{M}, w \models \otimes_i \varphi$ and suppose towards a contradiction that $\mathfrak{M}, w \not\models \otimes_i [i]\varphi$. Hence, $\mathfrak{M}, w \models \ominus_i \langle i \rangle \neg\varphi$. By semantic definition of $\ominus_i$ there is a $u \in \mathcal{R}_{\otimes_i}(w)$ such that $\mathfrak{M}, u \models \langle i \rangle \neg\varphi$ and by semantic definition of $\langle i \rangle$ we know there is a $v \in \mathcal{R}_{[i]}(u)$ such that $\mathfrak{M}, v \models \neg\varphi$. However, by **D4** we have $v \in \mathcal{R}_{\otimes_i}(w)$ too, and so $\mathfrak{M}, v \models \varphi$. Contradiction.

A15 Similar to A2.

A16 Similar to A3.

A17 Assume that $\mathfrak{M}, w \models \bigwedge_{i \in \mathsf{Agents}} [i]\varphi_i$. Hence, for each $i \in \mathsf{Agents}$ we have $\mathcal{R}_{[i]}(w) \subseteq \|\varphi_i\| = \{v \in W \mid \mathfrak{M}, v \models \varphi\}$. By straightforward semantic reasoning we obtain $\bigcap_{i \in \mathsf{Agents}} \mathcal{R}_{[i]}(w) \subseteq \bigcap_{i \in \mathsf{Agents}} \|\varphi_i\|$ and so $\bigcap_{i \in \mathsf{Agents}} \mathcal{R}_{[i]}(w) \subseteq \|\varphi_1 \wedge \ldots \wedge \varphi_n\|$. By **C6**, $\mathcal{R}_{[Ag]}(w) \subseteq \bigcap_{i \in \mathsf{Agents}} \mathcal{R}_{[i]}(w)$ and so $\mathcal{R}_{[Ag]}(w) \subseteq \|\varphi_1 \wedge \ldots \wedge \varphi_n\|$. Consequently, by the semantic definition of $[Ag]$, $\mathfrak{M}, w \models [Ag] \bigwedge_{i \in \mathsf{Agents}} \varphi_i$.

A19 Assume $\mathfrak{M}, w \models \mathsf{G}\varphi$ and suppose towards a contradiction that $\mathfrak{M}, w \not\models \mathsf{GG}\varphi$. Consequently, $\mathfrak{M}, w \models \mathsf{FF}\neg\varphi$. By semantic definition of F we know there is a $v \in \mathcal{R}_{\mathsf{G}}(w)$ and there is a $u \in \mathcal{R}_{\mathsf{G}}(v)$ such that $\mathfrak{M}, u \models \neg\varphi$. By **T1** we know $u \in \mathcal{R}_{\mathsf{G}}(w)$ and so $\mathfrak{M}, u \models \varphi$. Contradiction.

A20 Assume $\mathfrak{M}, w \models \mathsf{G}\varphi$. By **T1** there is a $v \in \mathcal{R}_{\mathsf{G}}(w)$ and thus by semantic definition of G we have $\mathfrak{M}, v \models \varphi$. Consequently, $\mathfrak{M}, w \models \mathsf{F}\varphi$.

A22 Assume $\mathfrak{M}, w \models \varphi$, by **T2** we know that for all $v \in \mathcal{R}_{\mathsf{G}}(w)$ there is a $u \in \mathcal{R}_{\mathsf{H}}(v)$ such that $u = w$. By semantic definition of P we thus have for all $v \in \mathcal{R}_{\mathsf{G}}(w)$, $\mathfrak{M}, v \models \mathsf{P}\varphi$. By semantic definition of G, we have $\mathfrak{M}, w \models \mathsf{GP}\varphi$.

A23 Similar to A22 using **T2**.

A24 Assume $\mathfrak{M}, w \models \mathsf{FP}\varphi$. Hence, there is a $v \in \mathcal{R}_{\mathsf{G}}(w)$ such that $Mo, v \models \mathsf{P}\varphi$ and there is a $u \in \mathcal{R}_{\mathsf{H}}(v)$ such that $\mathfrak{M}, u \models \varphi$. By **T2**, $w \in \mathcal{R}_{\mathsf{H}}(v)$. By **T3** we know that either (i) $u \in \mathcal{R}_{\mathsf{H}}(w)$, (ii) $u = w$, or (iii) $w \in \mathcal{R}_{\mathsf{H}}(u)$. We consider each case. Ad (i), then $\mathfrak{M}, w \models \mathsf{P}\varphi$. Ad (ii), then $\mathfrak{M}, w \models \varphi$. Ad (iii), then by **T2** $u \in \mathcal{R}_{\mathsf{G}}(w)$ and so $\mathfrak{M}, w \models \mathsf{F}\varphi$. Consequently, $\mathfrak{M}, w \models \mathsf{P}\varphi \vee \varphi \vee \mathsf{F}\varphi$.

A25 Similar to A24 using **T2** and **T4**.

A26 Assume $\mathfrak{M}, w \models \mathsf{F}\Diamond\varphi$. By semantic definition of F, there is a $v \in \mathcal{R}_{\mathsf{G}}(w)$ such that $\mathfrak{M}, v \models \Diamond\varphi$ and by semantic definition of $\Diamond$ there is a $u \in \mathcal{R}_{\Box}(v)$ with $\mathfrak{M}, u \models \varphi$. By **T5**, there is a $z \in \mathcal{R}_{[Ag]}(w)$ such that $u \in \mathcal{R}_{\mathsf{G}}(z)$. Consequently, $\mathfrak{M}, w \models \langle Ag \rangle \mathsf{F}\varphi$.

R3 Last, we show soundness of the irreflexivity-rule of TDS_n. Recall the rule:

$$\text{From } (\Box\neg p \wedge \Box(\mathsf{G}p \wedge \mathsf{H}p)) \rightarrow \varphi \text{ with } p \notin \mathsf{Atoms}(\varphi)\text{, infer } \varphi.$$

We assume that the atomic variable p does not occur in φ. We prove the result by contraposition and assume that φ is not TDS_n-valid. Therefore, we know there exists a TDS_n-model $\mathfrak{M} = \langle \mathfrak{F}, V \rangle$ s.t. $\mathfrak{F}$ is a TDS_n-frame and $\mathfrak{M}, w \not\models \varphi$ for some $w \in W$ of $\mathfrak{M}$. We define another TDS_n-model $\mathfrak{M}' = \langle \mathfrak{F}, V' \rangle$ over the frame $\mathfrak{F}$ and define the valuation V' as follows:

$$V'(q) := \begin{cases} V(q) & \text{if } q \neq p, \\ W \setminus \mathcal{R}_{\Box}(w) & \text{otherwise.} \end{cases}$$

(i.e., the valuation V' of p contains all worlds except those sharing the same moment with w). Clearly, since φ does not contain p and the other atomic propositions are evaluated in the same way in $\mathfrak{M}$ as in $\mathfrak{M}'$, we have $\mathfrak{M}', w \models \neg\varphi$. However, by the construction of V' and because $\mathfrak{F}$ is irreflexive by condition **T6**, we have that $\mathfrak{M}', w \models \Box\neg p \wedge \Box(\mathsf{G}p \wedge \mathsf{H}p))$. Since, $\mathfrak{M}', w \not\models \varphi$, by Definition 2.5, we have that $\mathfrak{M}', w \not\models (\Box\neg p \wedge \Box(\mathsf{G}p \wedge \mathsf{H}p)) \rightarrow \varphi$. Hence, we conclude that $(\Box\neg p \wedge \Box(\mathsf{G}p \wedge \mathsf{H}p)) \rightarrow \varphi$ is also not TDS_n-valid.

The above holds for each $i \in \mathsf{Agents}$, which finishes the proof of (†). We use (†) to prove the main claim. Assume $\Gamma \vdash_{\mathsf{TDS}_n} \varphi$. Then, by Definition 2.3 there exists a sequence $\varphi_1, ..., \varphi_n \in \mathcal{L}_n^{td}$ such that $\varphi_n = \varphi$, and for all $1 \leq i \leq n$, φ_i is (i) an TDS_n-theorem, (ii) an assumption from Γ, or (iii) a consequence of an application of R0 to some $\varphi_j = \psi$ and $\varphi_k = \psi \rightarrow \varphi_i$ with $j, k < i$. Take an arbitrary model $\mathfrak{M}$ and world w such that $\mathfrak{M}, w \models \Gamma$. By (†), for each $\varphi_i \in \Gamma$ for which (i) holds we have $\mathfrak{M}, w \models \varphi_i$. By assumption, for each φ_i for which (ii) holds, we have $\mathfrak{M}, w \models \varphi_i$. By validity of R0 and the previous two items, for each φ_i for which (iii) holds, we have $\mathfrak{M}, w \models \varphi_i$. Hence, $\mathfrak{M}, w \models \varphi$. QED

Due to the modularity of the above proof, we can see that soundness of the subsystem DS_n immediately follows from Theorem 2.1.

Corollary 2.1 (Soundness of DS_n)**.** *Let* DS_n *be the logic from Definition 2.2. For any formula* $\varphi \in \mathcal{L}_n^d$*, and any* $\Gamma \subseteq \mathcal{L}_n^d$*: if* $\Gamma \vdash_{\mathsf{DS}_n} \varphi$*, then* $\Gamma \models_{\mathsf{DS}_n} \varphi$*.*

2.2.1 Strong Completeness of Deontic STIT Logic

We first prove strong completeness for the atemporal deontic STIT logic DS_n. The results obtained in this section are also useful in proving completeness of TDS_n and the class of deontic STIT logics introduced in Chapter 3.

We adopt the *completeness via canonicity* method for normal modal logics (Blackburn et al., 2004). We prove the following claim:

$$\text{If } \Gamma \models_{\mathsf{DS}_n} \varphi \text{ then } \Gamma \vdash_{\mathsf{DS}_n} \varphi$$

for $\Gamma \subseteq \mathcal{L}^d_n$ and $\varphi \in \mathcal{L}^d_n$. The strategy is as follows: we define the notion of a DS_n-maximally consistent set of $\mathcal{L}^d_n$ formulae (Definition 2.6). These sets are used as worlds in constructing a canonical model for the logic DS_n (Definition 2.7). Subsequently, we prove a truth lemma (Lemma 2.5), ensuring that every DS_n-consistent set of formulae can be satisfied on the corresponding canonical model. The main aim is to demonstrate that the obtained canonical model is a DS_n-model (Theorem 2.6). Finally, the model is used to prove completeness via contraposition. Namely, if a formula φ is not DS_n-derivable from a set Γ, then $\{\neg\varphi\} \cup \Gamma$ is an DS_n-consistent set. By an adaptation of Lindenbaum's Lemma (Lemma 2.3) we know there is an DS_n-maximally consistent set Γ' extending $\{\neg\varphi\} \cup \Gamma$. Since Γ' is a world in the canonical DS_n-model, we know that $\neg\varphi$ and Γ are satisfiable and so $\Gamma \not\models_{\mathsf{DS}_n} \varphi$.

First, we define DS_n-consistent sets and DS_n-maximally consistent sets.

Definition 2.6 (DS_n-CS and DS_n-MCS)**.** *A set $\Delta \subset \mathcal{L}^d_n$ is a* DS_n-consistent set *(for short, DS_n-CS) iff $\Delta \nvdash_{\mathsf{DS}_n} \bot$. A set $\Delta \subset \mathcal{L}^d_n$ is a* DS_n-maximally consistent set *(for short, DS_n-MCS) iff Δ is a DS_n-CS and for any set $\Delta' \subseteq \mathcal{L}^d_n$ such that $\Delta \subset \Delta'$ it is the case that $\Delta' \vdash_{\mathsf{DS}_n} \bot$.*

We prove some useful properties of DS_n-MCSs, which are (implicitly) used throughout this section. In fact, the results hold for all modal logics considered in this thesis.

Lemma 2.2. *Let Γ be a MCS. Then, Γ has the following properties:*

- *$\Gamma \vdash_{\mathsf{DS}_n} \varphi$ iff $\varphi \in \Gamma$;*
- *$\varphi \in \Gamma$ iff $\neg\varphi \notin \Gamma$;*
- *$\varphi \wedge \psi \in \Gamma$ iff $\varphi \in \Gamma$ and $\psi \in \Gamma$.*

Proof. We prove each of the claims in turn:

(i) For the left-to-right direction assume that $\varphi \notin \Gamma$. Since Γ is a maximal, we know that $\Gamma \cup \{\varphi\}$ is inconsistent, i.e., $\Gamma \vdash_{\mathsf{DS}_n} \neg\varphi$. Due to the fact that Γ is consistent, we know that $\Gamma \nvdash_{\mathsf{DS}_n} \varphi$. For the opposite direction observe that if $\varphi \in \Gamma$, then trivially $\Gamma \vdash_{\mathsf{DS}_n} \varphi$.

(ii) Suppose that $\varphi \in \Gamma$. Observe that if $\neg\varphi \in \Gamma$ as well, then Γ would be inconsistent; hence, $\neg\varphi \notin \Gamma$. For the backward direction, assume that $\neg\varphi \notin \Gamma$. Suppose towards a contradiction that $\varphi \notin \Gamma$, then since Γ is a MCS, we know that both $\Gamma \cup \{\varphi\} \vdash_{\mathsf{TDS}_n} \bot$ and $\Gamma \cup \{\neg\varphi\} \vdash_{\mathsf{DS}_n} \bot$. However, this implies that $\Gamma \vdash_{\mathsf{DS}_n} \varphi \wedge \neg\varphi$, thus contradicting the consistency of Γ. Hence, we know that $\varphi \in \Gamma$.

(iii) If $\varphi \wedge \psi \in \Gamma$, then by fact (i) $\varphi \in \Gamma$ and $\psi \in \Gamma$ since both $\Gamma \vdash_{\mathsf{DS}_n} \varphi$ and $\Gamma \vdash_{\mathsf{DS}_n} \psi$ when $\varphi \wedge \psi \in \Gamma$. The opposite direction is proven similarly. QED

Adapting Lindenbaum's Lemma, every DS_n-CS can be extended to a DS_n-MCS.

Lemma 2.3 (Lindenbaum's Lemma for DS_n)**.** *Let $\Delta \subseteq \mathcal{L}_n^d$ be a DS_n-CS: there is a DS_n-MCS $\Delta' \subseteq \mathcal{L}_n^d$ such that $\Delta \subseteq \Delta'$.*

Proof. See (Blackburn et al., 2004, Lem. 4.17) for a general proof. QED

Definition 2.7 (Canonical model for DS_n)**.** *Let $[\alpha] \in \mathsf{Boxes} = \{\Box\} \cup \{[i] \mid i \in \mathsf{Agents}\} \cup \{\otimes_i \mid i \in \mathsf{Agents}\}$ and let $\langle\alpha\rangle$ be the operator dual to $[\alpha]$. We define the* canonical model *to be the tuple $\mathfrak{M}^c := \langle W^c, \mathcal{R}^c_\Box, \{\mathcal{R}^c_{[i]} \mid i \in \mathsf{Agents}\}, \{\mathcal{R}^c_{\otimes_i} \mid i \in \mathsf{Agents}\}, V^c\rangle$ such that:*

- *$W^c := \{\Gamma \subset \mathcal{L}_n^d \mid \Gamma$ is a DS_n-MCS$\}$;*
- *for each $[\alpha] \in \mathsf{Boxes}$ and each $\Delta \in W^c$, $\mathcal{R}^c_{[\alpha]}(\Delta) := \{\Gamma \in W^c \mid$ for all $[\alpha]\varphi \in \Delta$, $\varphi \in \Gamma\}$;*
- *V^c is a valuation function such that for all $p \in \mathsf{Atoms}$, $V^c(p) := \{\Delta \in W^c \mid p \in \Delta\}$.*

The semantic evaluation of formulae from $\mathcal{L}_n^d$ is defined as usual (Definition 2.5).

We show some useful properties for demonstrating that the canonical model is a DS_n-model (Lemma 2.6).

Lemma 2.4 (Existence Lemma)**.** *Let $[\alpha] \in \mathsf{Boxes}$ and let $\langle\alpha\rangle$ be the operator dual to $[\alpha]$.*[8] *For any world $\Delta \in W^c$ of $\mathfrak{M}^c$ and each $i \in \mathsf{Agents}$ the following holds:*

- *If $\langle\alpha\rangle\varphi \in \Delta$, then there is a $\Gamma \in W^c$ such that $\varphi \in \Gamma$ and $\Gamma \in \mathcal{R}^c_{\alpha]}(\Delta)$.*

Proof. See (Blackburn et al., 2004, Lem. 4.20) for a general proof. QED

Corollary 2.2. *Let $[\alpha] \in \mathsf{Boxes}$ and let $\langle\alpha\rangle$ be the operator dual to $[\alpha]$. For any world $\Delta \in W^c$ of $\mathfrak{M}^c$ and each $i \in \mathsf{Agents}$ the following holds:*

- *If for all $\Gamma \in \mathcal{R}^c_{[\alpha]}(\Delta), \varphi \in \Gamma$, then $[\alpha]\varphi \in \Delta$.*

The following lemma shows that the defined model is canonical for DS_n, i.e., each DS_n-MCS is satisfiable on this model.

Lemma 2.5 (Truth Lemma)**.** *For any $\varphi \in \mathcal{L}_n^d$ and $\Delta \in W^c$ of $\mathfrak{M}^c$: $\mathfrak{M}^c, \Delta \models \varphi$ iff $\varphi \in \Delta$.*

[8] The diamond-shaped operator is the defined dual of its box-shaped counterpart. Consequently, in the syntactical construction of the canonical model $\mathfrak{M}^c$, when we write $\Diamond$, we denote the syntactic object $\neg\Box\neg$. For readability, we use the defined operator $\Diamond$.

Proof. The proof is by induction on the complexity of φ. *Base case* $\varphi = p$. Follows directly from the definition of V^c in Definition 2.7. *Inductive Step.* The cases for the propositional connectives $\neg$ and $\wedge$ are straightforward, see (Blackburn et al., 2004, Lem.4.21). We show the case for the modality $[\alpha] \in \mathsf{Boxes}$:

($\varphi = [\alpha]\psi$) **Left-to-Right.** Suppose $\mathfrak{M}^c, \Delta \models [\alpha]\psi$, then for all $\Gamma \in \mathcal{R}^c_{[\alpha]}(\Delta)$, $\mathfrak{M}^c, \Gamma \models \psi$. By IH, for all $\Gamma \in \mathcal{R}^c_{[\alpha]}(\Delta)$, $\psi \in \Gamma$. By Corollary 2.2, $[\alpha]\psi \in \Delta$.

Right-to-Left. Suppose $[\alpha]\psi \in \Delta$, and take an arbitrary $\Gamma \in \mathcal{R}^c_{[\alpha]}(\Delta)$, then by definition of $\mathcal{R}^c_{[\alpha]}$ we have $\psi \in \Gamma$. By IH, $\mathfrak{M}^c, \Gamma \models \psi$ and since Γ was arbitrary by semantic definition of $[\alpha]$ we conclude $\mathfrak{M}^c, \Delta \models [\alpha]\psi$. QED

Lemma 2.6 (Canonical DS_n-model)**.** *The canonical model $\mathfrak{M}^c$ is a DS_n-model.*

Proof. W^c and V^c are trivially well-defined. We only need to show that $\mathfrak{M}^c$ satisfies the properties **C1**–**C4** and **D1**–**D4** of Definition 2.4. Take an arbitrary $\Delta \in W^c$ of $\mathfrak{M}^c$:

C1 To prove that $\mathcal{R}^c_\Box$ is an equivalence relation, it suffices to show that $\mathcal{R}^c_\Box$ is (i) reflexive and (ii) euclidean. Ad (i), take an arbitrary $\varphi \in \mathcal{L}^d_n$ and assume $\Box\varphi \in \Delta$. Since Δ is a DS_n-MCS we know that $\Box\varphi \to \varphi \in \Delta$ (axiom $\mathsf{A2}$). Consequently, $\varphi \in \Delta$. Since φ was arbitrary we know by Definition 2.7 that $\Delta \in \mathcal{R}^c_\Box(\Delta)$. Ad (ii), assume that $\Gamma, \Sigma \in \mathcal{R}^c_\Box(\Delta)$. We show that $\Gamma \in \mathcal{R}^c_\Box(\Sigma)$. Take an arbitrary $\varphi \in \mathcal{L}^d_n$ and assume $\Box\varphi \in \Sigma$. Suppose towards a contradiction that $\varphi \notin \Gamma$. Consequently, since $\Gamma \in \mathcal{R}^c_\Box(\Delta)$, $\Diamond\neg\varphi \in \Delta$. By the fact that Δ is a DS_n-MCS we know $\Diamond\neg\varphi \to \Box\Diamond\neg\varphi \in \Delta$ (axiom $\mathsf{A3}$) and so $\Box\Diamond\neg\varphi \in \Delta$. By the assumption that $\Sigma \in \mathcal{R}^c_\Box(\Delta)$ we have $\Diamond\neg\varphi \in \Sigma$ and so $\neg\Box\varphi \in \Sigma$. This contradicts the assumption that Σ is a DS_n-MCS.

C2 Similar to **C1**.

C3 Consider an arbitrary $\Gamma \in \mathcal{R}^c_{[i]}(\Delta)$. We prove that $\Gamma \in \mathcal{R}^c_\Box(\Delta)$. Take an arbitrary $\varphi \in \mathcal{L}^d_n$ and assume that $\Box\varphi \in \Delta$. By the fact that Δ is a DS_n-MCS, we know that $\Box\varphi \to [i]\varphi \in \Delta$ (axiom $\mathsf{A7}$). Consequently, $[i]\varphi \in \Delta$ and thus $\varphi \in \Gamma$. Since φ was arbitrary we know by Definition 2.7 that $\Gamma \in \mathcal{R}^c_\Box(\Delta)$.

C4 Let $\Gamma_1, ..., \Gamma_n \in \mathcal{R}^c_\Box(\Delta)$. We show that there is a $\Sigma \in W^c$ such that $\Sigma \in \bigcap_{i \in \mathsf{Agents}} \mathcal{R}^c_{[i]}(\Gamma_i)$. We construct this DS_n-MCS Σ. Consider the following set:

$$\Sigma' = \bigcup_{i \in \mathsf{Agents}} \{\varphi \mid [i]\varphi \in \Gamma_i\} \cup \{\psi \mid \Box\psi \in \Delta\}$$

We suppose towards a contradiction that Σ' is inconsistent, i.e., $\Sigma' \vdash_{\mathsf{DS}_n} \bot$. Consequently, we know that there are $\varphi_1, ..., \varphi_k \in \bigcup_{i \in \mathsf{Agents}} \{\varphi \mid [i]\varphi \in \Gamma_i\}$ and there are $\psi_1, ..., \psi_l \in \{\psi \mid \Box\psi \in \Delta\}$ such that

$$(\dagger) \quad \vdash_{\mathsf{DS}_n} (\varphi_1 \wedge ... \wedge \varphi_k) \to (\neg\psi_1 \vee ... \vee \neg\psi_l)$$

We define for each $i \in \mathsf{Agents}$ the following set $\Phi_i = \{\varphi_m \mid [i]\varphi_m \in \Gamma_i\} \cap \{\varphi_1, ..., \varphi_k\}$. As an immediate consequence, we have $\bigwedge [i]\Phi_i \in \Gamma_i$, and since $[i]$ is a normal modal operator $[i] \bigwedge \Phi_i \in \Gamma_i$. Since $\Gamma_i \in \mathcal{R}^c_\Box(\Delta)$ we have $\Diamond [i] \bigwedge \Phi_i \in \Delta$ for each $i \in \mathsf{Agents}$ and consequently $\bigwedge_{i\in\mathsf{Agents}} \Diamond[i]\bigwedge\Phi_i \in \Delta$. Since Δ is a DS_n-MCS, we know that $\bigwedge_{i\in\mathsf{Agents}}\Diamond[i]\bigwedge\Phi_i \to \Diamond\bigwedge_{i\in\mathsf{Agents}}[i]\bigwedge\Phi_i \in \Delta$ (axiom $\mathsf{A8}$) and so $\Diamond\bigwedge_{i\in\mathsf{Agents}}[i]\bigwedge\Phi_i \in \Delta$. By the existence lemma 2.4, there is a $\Sigma \in \mathcal{R}^c_\Box(\Delta)$ such that $\bigwedge_{i\in\mathsf{Agents}}[i]\bigwedge\Phi_i \in \Sigma$. By the fact that $[i]\bigwedge\Phi_i \to \bigwedge\Phi_i \in \Sigma$ (axiom $\mathsf{A5}$), we have $\bigwedge_{i\in\mathsf{Agents}}\bigwedge\Phi_i\in\Sigma$. By (†) we have $\neg\psi_1\vee...\vee\neg\psi_l \in \Sigma$ but since $\Sigma\in\mathcal{R}^c_\Box(\Delta)$ we also have $\psi_1,...,\psi_l\in\Sigma$. Contradiction. Hence, Σ' is DS_n-consistent. By Lemma 2.3 we know there is a DS_n-MCS $\Sigma \in W^c$ extending Σ'. Last, by the construction of $\Sigma' \subseteq \Sigma$ and the definitions of $\mathcal{R}^c_\Box$ and $\mathcal{R}^c_{[i]}$ we have $\Sigma\in\mathcal{R}^c_\Box(\Delta)$ and $\Sigma \in \mathcal{R}^c_{[i]}(\Gamma_i)$ for each $i\in\mathsf{Agents}$.

D1 Suppose that $\Gamma \in \mathcal{R}^c_{\otimes_i}(\Delta)$ and $\Sigma \in \mathcal{R}^c_\Box(\Delta)$. We prove that $\Gamma \in \mathcal{R}^c_{\otimes_i}(\Sigma)$. Take an arbitrary $\varphi \in \mathcal{L}^d_n$ and suppose that $\otimes_i\varphi \in \Sigma$. Since Σ is a DS_n-MCS we know $\otimes_i\varphi \to \Box \otimes_i \varphi \in \Sigma$ (axiom $\mathsf{A10}$) and thus $\Box \otimes_i \varphi \in \Sigma$. By the fact that $\mathcal{R}^c_\Box$ is an equivalence class (see **C1** above) we know $\Delta \in \mathcal{R}^c_\Box(\Sigma)$ and so $\otimes_i\varphi \in \Delta$. By the assumption that $\Gamma \in \mathcal{R}^c_{\otimes_i}(\Delta)$ we have $\varphi \in \Gamma$. Since φ was arbitrary, by Definition 2.7, we know that $\Gamma \in \mathcal{R}^c_{\otimes_i}(\Sigma)$.

D2 We show that there is a Γ such that $\Gamma \in \mathcal{R}^c_\Box(\Delta)$ and for all $\Sigma \in \mathcal{R}^c_{[i]}(\Gamma)$ we have $\Sigma \in \mathcal{R}^c_{\otimes_i}(\Delta)$. We construct Γ. Let

$$\Gamma' = \{[i]\varphi \mid \otimes_i \varphi \in \Delta\} \cup \{\psi \mid \Box\psi \in \Delta\}$$

Suppose towards a contradiction that Γ' is inconsistent. Then we know that

$$\vdash_{\mathsf{DS}_n} ([i]\varphi_i \wedge ... \wedge [i]\varphi_k \wedge \psi_1 \wedge ... \wedge \psi_l) \to \bot$$

where $[i]\varphi_1, ..., [i]\varphi_k \in \{[i]\varphi \mid \otimes_i \varphi \in \Gamma\}$ and $\psi_1, ..., \psi_l \in \{\psi \mid \Box\psi \in \Gamma\}$. Let $\hat{\varphi} = \varphi_1 \wedge ... \wedge \varphi_k$ and $\hat{\psi} = \psi_1 \wedge ... \wedge \psi_l$. By normality of $[i]$, we have $\vdash_{\mathsf{DS}_n} [i]\hat{\varphi} \equiv ([i]\varphi_1 \wedge ... \wedge [i]\varphi_k)$ and thus by basic modal reasoning we obtain $\vdash_{\mathsf{DS}_n} \hat{\psi} \to \neg[i]\hat{\varphi}$. By the normality of $\Box$, we have $\vdash_{\mathsf{DS}_n} \Box\hat{\psi} \to \Box\neg[i]\hat{\varphi}$, which implies $\vdash_{\mathsf{DS}_n} \Box\hat{\psi} \to \neg\Diamond[i]\hat{\varphi}$. Clearly, because $\Box\hat{\psi} \in \Delta$ and the fact that Δ is a DS_n-MCS, we know that $\neg\Diamond[i]\hat{\varphi} \in \Delta$. Also, since $\otimes_i\varphi_1, ..., \otimes_i\varphi_k \in \Delta$ and $\otimes_i$ is a normal modal operator, we have that $\otimes_i\hat{\varphi} \in \Delta$ as well. We know that $\otimes_i\hat{\varphi} \to \Diamond[i]\hat{\varphi} \in \Delta$ (axiom $\mathsf{A11}$) and thus by the fact that Δ is a DS_n-MCS, we obtain $\Diamond[i]\hat{\varphi} \in \Delta$. We have a contradiction, and so Γ' is consistent. By Lemma 2.3 there is a DS_n-MCS Γ such that $\Gamma' \subseteq \Gamma$. By the definition of $\mathcal{R}^c_\Box$ and the construction of $\Gamma' \subseteq \Gamma$ we know $\Gamma \in \mathcal{R}^c_\Box(\Delta)$. Last, assume $\Sigma \in \mathcal{R}^c_{[i]}(\Gamma)$ and take an arbitrary $\psi \in \mathcal{L}^d_n$ with $\otimes_i\psi \in \Delta$. By construction of Γ, $[i]\psi \in \Gamma$ and thus $\psi \in \Sigma$. Since ψ was arbitrary we have by the definition of $\mathcal{R}^c_{\otimes_i}$ that $\Sigma \in \mathcal{R}^c_{\otimes_i}(\Delta)$.

D3 Similar to **C3**.

D4 Suppose that $\Gamma \in \mathcal{R}^c_{\otimes_i}(\Delta)$ and $\Sigma \in \mathcal{R}^c_{[i]}(\Gamma)$. We prove that $\Sigma \in \mathcal{R}^c_{\otimes_i}(\Delta)$. Take an arbitrary $\varphi \in \mathcal{L}^d_n$ and assume $\otimes_i\varphi \in \Delta$. Since Δ is a DS_n-MCS, we know $\otimes_i\varphi \rightarrow \otimes_i[i]\varphi \in \Delta$ (axiom A13). Consequently, $\otimes_i[i]\varphi \in \Delta$. Since $\Gamma \in \mathcal{R}^c_{\otimes_i}(\Delta)$ we have $[i]\varphi \in \Gamma$ and since $\Sigma \in \mathcal{R}^c_{[i]}(\Gamma)$ we have $\varphi \in \Sigma$. Last, because φ was arbitrary we have, by Lemma 2.7, that $\Sigma \in \mathcal{R}^c_{\otimes_i}(\Delta)$. QED

We can now demonstrate strong completeness of DS_n.

Theorem 2.2 (Strong Completeness of DS_n)**.** *For any formula $\varphi \in \mathcal{L}^d_n$, and any $\Gamma \subseteq \mathcal{L}^d_n$: if $\Gamma \models_{\mathsf{DS}_n} \varphi$, then $\Gamma \vdash_{\mathsf{DS}_n} \varphi$.*

Proof. The proof is by contraposition. Suppose φ is not DS_n-derivable from Γ. This means that $\Gamma \cup \{\neg\varphi\}$ is a DS_n-CS. Namely, if $\Gamma \cup \{\neg\varphi\}$ would be DS_n-inconsistent, then $\Gamma, \neg\varphi \vdash_{\mathsf{DS}_n} \bot$ and so $\Gamma \vdash_{\mathsf{DS}_n} \varphi$. By Lemma 2.3 there is a $\Gamma' \subseteq \mathcal{L}^d_n$ such that Γ' is a DS_n-MCS and $\Gamma \cup \{\neg\varphi\} \subseteq \Gamma'$. By construction of the canonical model, $\Gamma' \in W^c$ and by Lemma 2.5 we know that $\mathfrak{M}^c, \Gamma' \models \Gamma$ and $\mathfrak{M}^c, \Gamma' \models \neg\varphi$. By Lemma 2.6, $\mathfrak{M}^c$ is a DS_n-model and so $\Gamma \not\models_{\mathsf{DS}_n} \varphi$. QED

2.2.2 Weak Completeness of Temporal Deontic STIT Logic

Following Gabbay et al. (1994), due to the use of the irreflexivity rule R3, we cannot readily adapt the standard *completeness via canonicity* method for normal modal logics (Blackburn et al., 2004). In order to prove completeness of the logic TDS_n, we must define a specific canonical model, i.e., one that respects temporal irreflexivity. To ensure irreflexivity, we adopt the mechanism from Gabbay et al. (1994) and employed by Lorini (2013) in the context of STIT, which allows us to encode TDS_n-MCSs with information that excludes reflexive points in the resulting model.

The strategy is as follows: we define the notion of a TDS_n-maximally consistent set (MCS) of $\mathcal{L}^{td}_n$ formulae (Definition 2.8). These MCSs are used as worlds in constructing a canonical model for the logic TDS_n (Definition 2.9). Subsequently, we define a specific submodel of the canonical model, restricted to specific TDS_n-MCSs called IRR-theories. An IRR-theory is a TDS_n-MCS constructed in such a way that it coherently and uniquely labels itself and each reachable TDS_n-MCS. The labeling occurs by giving each MCS a unique atomic proposition that identifies it (Definition 2.12). The truth lemma (Lemma 2.7) holds for the canonical submodel restricted to IRR theories. It is then shown that each TDS_n-consistent formula $\varphi \in \mathcal{L}^{td}_n$ can be consistently extended to an IRR theory (Lemma 2.8). The gist of the proof lies in the observation that one needs infinitely many atomic propositions to coherently name each reachable world in an infinite branching time structure. Since a formula φ contains only finitely many atoms, and since the set Atoms is infinite, there are infinitely many atoms left to coherently define the IRR theory extending $\{\varphi\}$. The labeling strategy is used to demonstrate that the obtained canonical submodel is a TDS_n-model (Theorem 2.11). As a last step, this TDS_n-model is used to prove weak completeness via contraposition in the usual way.

It must be noted that we only obtain weak completeness of TDS_n, e.g., see (Gabbay et al., 1994). In order to prove strong completeness, we must guarantee that any arbitrary TDS_n-consistent set Δ can be extended to an IRR theory. However, to ensure that a TDS_n-maximally consistent extension Δ' of Δ is an IRR theory, we need an infinite number of atomic formulae p not occurring in Δ. Hence, Δ' cannot contain infinitely many atomic formulae, e.g., when $\Delta' = \mathsf{Atoms}$. This observation excludes TDS_n-consistent sets that are maximal but are not IRR theories. Consequently, not every arbitrary TDS_n-consistent set can be extended to an IRR theory. As observed, we can still guarantee weak completeness because every TDS_n-consistent formula φ is syntactically finite, which means that there is an infinite number of atoms in Atoms not occurring in φ.

We now turn to the proof. First, we define TDS_n-maximally consistent sets and the general canonical model for TDS_n that does not yet ensure irreflexivity. The definitions are similar to the canonical model construction for the logic DS_n.

Definition 2.8 (TDS_n-CS and TDS_n-MCS)**.** *A set $\Delta \subset \mathcal{L}_n^{td}$ is an TDS_n* consistent set *(for short, TDS_n-CS) iff $\Delta \nvdash_{\mathsf{TDS}_n} \bot$. A set $\Delta \subset \mathcal{L}_n^{td}$ is an TDS_n*-maximally consistent set *(for short, TDS_n-MCS) iff Δ is an TDS_n-CS and for any set $\Delta' \subseteq \mathcal{L}_n^{td}$ such that $\Delta \subset \Delta'$ it is the case that $\Delta' \vdash_{\mathsf{TDS}_n} \bot$.*

Observe that the properties of MCSs proven in Lemma 2.2 also hold for TDS_n-MCSs. For the remainder of this section, we use Boxes to refer to the set of box-shaped modalities of $\mathcal{L}_n^{td}$, i.e., $\mathsf{Boxes} := \{\Box, [Ag], \mathsf{G}, \mathsf{H}\} \cup \{[i] \mid i \in \mathsf{Agents}\} \cup \{\otimes_i \mid i \in \mathsf{Agents}\}$.

Definition 2.9 (Canonical model for TDS_n)**.** *Let $[\alpha] \in \mathsf{Boxes}$ and let $\langle\alpha\rangle$ be the operator dual to $[\alpha]$. We define the* canonical model *to be the tuple $\mathfrak{M}^c := \langle W^c, \mathcal{R}^c_\Box, \{\mathcal{R}^c_{[i]} \mid i \in \mathsf{Agents}\}, \{\mathcal{R}^c_{\otimes_i} \mid i \in \mathsf{Agents}\}, \mathcal{R}^c_{[Ag]}, \mathcal{R}^c_{\mathsf{G}}, \mathcal{R}^c_{\mathsf{H}}, V^c\rangle$ such that:*

- *$W^c := \{\Gamma \subset \mathcal{L}_n^{td} \mid \Gamma$ is a TDS_n-MCS$\}$;*
- *for each $[\alpha] \in \mathsf{Boxes}$ and for all $\Delta \in W^c$, $\mathcal{R}^c_{[\alpha]}(\Delta) := \{\Gamma \in W^c \mid$ for all $[\alpha]\varphi \in \Delta$, then $\varphi \in \Gamma\}$;*
- *V^c is a valuation function such that for all $p \in \mathsf{Atoms}$, $V^c(p) := \{\Delta \in W^c \mid p \in \Delta\}$.*

We are interested in a submodel of the canonical model, namely, one that excludes reflexive worlds. In order to guarantee that the submodel satisfies the truth lemma (Lemma 2.7) we must ensure that the submodel is well-defined. For this, we adopt Lorini's (2013) notion of a diamond-saturated set.

Definition 2.10 (Diamond-saturated set (Lorini, 2013))**.** *Let X be a set of MCSs and let $\langle\alpha\rangle$ be dual to $[\alpha] \in \mathsf{Boxes}$. We say that X is a* diamond saturated set *iff for all $\Gamma \in X$, for each $\langle\alpha\rangle\varphi \in \Gamma$ there exists a $\Delta \in X$ such that $\mathcal{R}^c_{[\alpha]}\Gamma\Delta$ and $\varphi \in \Delta$.*

Definition 2.11 (An X-induced submodel $\mathfrak{M}^X$)**.** *Let* $\mathfrak{M}^c := \langle W^c, \mathcal{R}^c_{\Box}, \{\mathcal{R}^c_{[i]} \mid i \in \mathsf{Agents}\}, \{\mathcal{R}^c_{\otimes_i} \mid i \in \mathsf{Agents}\}, \mathcal{R}^c_{[Ag]}, \mathcal{R}^c_{\mathsf{G}}, \mathcal{R}^c_{\mathsf{H}}, V^c\rangle$ *be the canonical model from Definition 2.9. Let* $X \subseteq W^c$*. We define the* X *induced submodel* $\mathfrak{M}^X = \langle W^X, \mathcal{R}^X_{\Box}, \{\mathcal{R}^X_{[i]} \mid i \in \mathsf{Agents}\}, \{\mathcal{R}^X_{\otimes_i} \mid i \in \mathsf{Agents}\}, \mathcal{R}^X_{[Ag]}, \mathcal{R}^X_{\mathsf{G}}, \mathcal{R}^X_{\mathsf{H}}, V^X\rangle$ *of* $\mathfrak{M}^c$ *as follows:*

- $W^X := W^c \cap X$*;*
- *For each* $[\alpha] \in \mathsf{Boxes}$, $\mathcal{R}_{[\alpha]}|_X := \{(\Gamma, \Delta) \mid (\Gamma, \Delta) \in \mathcal{R}_{[\alpha]}$ *and* $\Gamma, \Delta \in X\}$*;*
- *For each* $p \in \mathsf{Atoms}$, $V^X(p) := V^c(p) \cap X$.

Lemma 2.7 (Truth Lemma)**.** *Let* $\mathfrak{M}^c$ *be the canonical model and let* $X \subseteq W^c$ *be a diamond saturated set with* $\Gamma \in X$, $\varphi \in \mathcal{L}^{td}_n$*. Let* $\mathfrak{M}^X$ *be the* X *induced submodel of* $\mathfrak{M}^c$*. Then,* $\mathfrak{M}^X, \Gamma \models \varphi$ *iff* $\varphi \in \Gamma$*.*

Proof. Proven in the usual manner (Blackburn et al., 2004, Lem. 4.70). QED

Following Lorini (2013), let IRR-theories be those sets of TDS_n-formulae that (i) are maximally consistent, (ii) contain a label $name(p) := \Box\neg p \wedge \Box(\mathsf{G}p \wedge \mathsf{H}p)$, uniquely labeling a *moment* and (iii) for any world that is reachable through any 'zig-zagging' sequence of diamond operators, that is, every zig-zagging formula φ of the form

$$\langle\alpha_1\rangle(\varphi_1 \wedge \langle\alpha_2\rangle(\varphi_2 \wedge ... \wedge \langle\alpha_n\rangle\varphi_n))...)$$

where $\langle\alpha_i\rangle$ is dual to $[\alpha_i] \in \mathsf{Boxes}$ with $1 \leq i \leq n$, there exists a corresponding zig-zagging formula $\varphi(q)$ (where q is a propositional variable) of the form,

$$\langle\alpha_1\rangle(\varphi_1 \wedge \langle\alpha_2\rangle(\varphi_2 \wedge ... \wedge \langle\alpha_n\rangle(\varphi_n \wedge \Box\neg q \wedge \Box(\mathsf{G}q \wedge \mathsf{H}q)))...)$$

labeling reachable worlds.

Intuitively, the naming formula $\Box\neg p \wedge \Box(\mathsf{G}p \wedge \mathsf{H}p) \in \Gamma$ ensures that the literal $\neg p$ uniquely identifies the moment of which Γ is part in the constructed canonical model (i.e., $\Box\neg p$). It is unique because all other moments making up the tree-structure of which Γ is part will satisfy p instead (i.e., $\Box(\mathsf{G}p \wedge \mathsf{H}p)$). The inclusion of zig-zagging formulae ensures that any other moment in the desired branching time structure, reachable through sequences of diamond operators, will likewise be uniquely named. Subsequently, using naming formulae (first item of Definition 2.12) and zig-zagging formulae (second item of Definition 2.12) in the selection of TDS_n-MCSs enables us to ensure that each moment in the canonical model is irreflexive (Gabbay et al., 1994).

Definition 2.12 (IRR-theory (Lorini, 2013))**.** *Let* Zig *be the set of all zig-zagging formulae in* $\mathcal{L}^{td}_n$ *and let* name(p):= $\Box\neg p \wedge \Box(\mathsf{G}p \wedge \mathsf{H}p)$ *where* p *is a propositional variable. A set of formulae* Γ *is called an* IRR-theory *iff the following hold:*

- *Γ is a TDS_n-MCS and name$(p) \in \Gamma$, for some propositional variable p;*
- *if $\varphi \in \Gamma \cap \mathsf{Zig}$, then $\varphi(q) \in \Gamma$, for some propositional variable q.*

Let $\mathsf{IRR} := \{\Gamma \subseteq \mathcal{L}_n^{td} \mid \Gamma$ is an IRR-theory $\}$ denote the set of all IRR-theories.

The proof of the following lemma demonstrates how for each TDS_n-consistent formula $\varphi \in \mathcal{L}_n^{td}$, we can construct an IRR-theory containing it.

Lemma 2.8. *Let $\varphi \in \mathcal{L}_n^{td}$ be a consistent formula. Then, there exists an IRR-theory Γ such that $\varphi \in \Gamma$.*

Proof. Let $\varphi \in \mathcal{L}_n^{td}$ be a consistent formula. We enumerate the formulae of $\mathcal{L}_n^{td}$ so that each formula in odd position is an element of Zig and make use of this enumeration to build an increasing sequence of consistent theories $\Gamma_0, \Gamma_1, \ldots, \Gamma_n, \ldots$

We let $\Gamma_0 := \{\varphi \wedge \Box\neg p \wedge \Box(\mathsf{G}p \wedge \mathsf{H}p)\}$ for some propositional variable p not occurring in φ. We define the sequence of Γ_n (for $n > 0$) as follows: Assume that Γ_n is defined and consider ψ_n of the enumeration. We know that either $\Gamma_n \cup \{\neg\psi_n\}$ is consistent or $\Gamma_n \cup \{\psi_n\}$ is consistent. If $\Gamma_n \cup \{\neg\psi_n\}$ is consistent, set $\Gamma_{n+1} := \Gamma_n \cup \{\neg\psi_n\}$. If $\Gamma_n \cup \{\psi_n\}$ is consistent, then there are two cases to consider: either n is even, or n is odd. If n is even, then set $\Gamma_{n+1} := \Gamma_n \cup \{\psi_n\}$. Otherwise, if n is odd, set $\Gamma_{n+1} := \Gamma_n \cup \{\psi_n, \psi_n(q)\}$, where q is a propositional variable not occurring in Γ_n or ψ. We define our desired maximally consistent IRR-theory as follows:

$$\Gamma := \bigcup_{n \in \mathbb{N}} \Gamma_n$$

To finish the proof, we need to show that Γ is both a TDS_n-MCS and an IRR-theory. We first prove that (i) Γ is a MCS and then show that (ii) Γ is an IRR-theory.

To prove claim (i), it is useful to first prove that for all $n \in \mathbb{N}$, each Γ_n is consistent. We show this claim by induction on n. In the base case, assume for a contradiction that $\Gamma_0 = \{\varphi \wedge \Box\neg p \wedge \Box(\mathsf{G}p \wedge \mathsf{H}p)\}$ is inconsistent. Hence, $\Box\neg p \wedge \Box(\mathsf{G}p \wedge \mathsf{H}p) \wedge \varphi \vdash_{\mathsf{TDS}_n} \bot$, which further implies that $\vdash_{\mathsf{TDS}_n} \Box\neg p \wedge \Box(\mathsf{G}p \wedge \mathsf{H}p) \rightarrow (\varphi \rightarrow \bot)$. We may infer from the rule $\mathsf{R3}$ that $\vdash_{\mathsf{TDS}_n} \varphi \rightarrow \bot$. However, we know that φ is consistent, meaning that $\not\vdash_{\mathsf{TDS}_n} \varphi \rightarrow \bot$. We have thus obtained a contradiction implying that Γ_0 is consistent. For the inductive step, assume that Γ_n is consistent. We want to show that Γ_{n+1} is consistent. This trivially follows by the definition of Γ_{n+1}.

To prove that Γ is a MCS, we must show that Γ is both consistent and maximal. Assume for a contradiction that Γ is inconsistent. Then, this implies that for some finite subset Γ' of Γ, $\Gamma' \vdash \bot$. However, if this is the case, then there exists some Γ_n such that $\Gamma_n \vdash_{\mathsf{TDS}_n} \bot$. We know this cannot be the case by the previous paragraph, and so, Γ must be consistent. Assume now that there exists some Γ' such that $\Gamma \subset \Gamma'$ and $\Gamma' \not\vdash_{\mathsf{TDS}_n} \bot$. Let $\psi \in \Gamma' \setminus \Gamma$. Since ψ is a formula in $\mathcal{L}_n^{td}$, we know that if was considered at some point during the construction of the sequence $\Gamma_0, \Gamma_1, \ldots, \Gamma_n, \ldots$. Since $\psi \notin \Gamma$ this implies that there exists some Γ_m such that $\Gamma_m \cup \{\psi\}$ is inconsistent. Therefore, $\Gamma_m \vdash_{\mathsf{TDS}_n} \neg\psi$, which implies that $\Gamma \vdash_{\mathsf{TDS}_n} \neg\psi$. Due to the fact that $\Gamma \subset \Gamma'$, it follows that $\Gamma' \vdash_{\mathsf{TDS}_n} \neg\psi$ and $\Gamma' \vdash_{\mathsf{TDS}_n} \psi$ since $\psi \in \Gamma'$, which is a contradiction. Therefore, Γ is a MCS.

We now prove that Γ is an IRR-theory. By construction, we know that $\varphi \wedge \Box\neg p \wedge \Box(\mathsf{G}p \wedge \mathsf{H}p) \in \Gamma_0 \subset \Gamma$, and since Γ is a MCS, it follows that $\Box\neg p \wedge \Box(\mathsf{G}p \wedge \mathsf{H}p) \in \Gamma$, thus satisfying the first condition of being an IRR-theory. The second condition of being an IRR-theory is satisfied by the fact that whenever a formula $\psi \in \mathsf{Zig}$ is added to $\Gamma_m \subset \Gamma$, for $m \in \mathbb{N}$, the formula $\psi(q)$ is added as well with q fresh. QED

The following existence lemma guarantees that the set IRR is a *diamond saturated* set (Definition 2.10), which implies that the submodel $\mathfrak{M}^{\mathsf{IRR}}$ obtained by restricting the canonical model to IRR theories satisfies the truth lemma (Lemma 2.7).

Lemma 2.9 (Existence lemma)**.** *Let $\Gamma \in \mathsf{IRR}$ be an IRR-theory and let $\langle\alpha\rangle$ be dual to $[\alpha] \in \mathsf{Boxes}$. For each $\langle\alpha\rangle\varphi \in \Gamma$ there exists an IRR-theory $\Delta \in \mathsf{IRR}$ such that $\Delta \in \mathcal{R}^c_{[\alpha]}(\Gamma)$ and $\varphi \in \Delta$.*

Proof. The proof is the same as in Lorini (2013, Lem. 16). QED

Henceforth, we use the superscript IRR for denoting the elements of the IRR induced canonical submodel $\mathfrak{M}^{\mathsf{IRR}}$. We prove the following useful lemma:

Lemma 2.10. *Let $\mathfrak{M}^{\mathsf{IRR}}$ be the IRR induced submodel of $\mathfrak{M}^c$. Let $\langle\alpha\rangle$ be dual to $[\alpha] \in \mathsf{Boxes}$ and let $\Gamma, \Delta \in \mathsf{IRR}$. Then, $\Delta \in \mathcal{R}^{\mathsf{IRR}}_{[\alpha]}(\Gamma)$ iff for all $\varphi \in \Delta$, $\langle\alpha\rangle\varphi \in \Gamma$.*

Proof. **Left-to-Right.** Assume $\Delta \in \mathcal{R}^{\mathsf{IRR}}_{[\alpha]}(\Gamma)$. By Definition 2.11 we know that $\Delta \in \mathcal{R}^c_{[\alpha]}(\Gamma)$. By the definition of $\mathcal{R}^c_{[\alpha]}$ (Definition 2.9) we know that for all $[\alpha]\varphi \in \Gamma$, $\varphi \in \Delta$. Which by contraposition gives us for all $\varphi \in \Delta$, $\langle\alpha\rangle\varphi \in \Gamma$.

Right-to-Left. Assume that for all $\varphi \in \mathcal{L}^{td}_n$, if $\varphi \in \Delta$, then $\langle\alpha\rangle\varphi \in \Gamma$. Take an arbitrary $\psi \in \mathcal{L}^{td}_n$ and suppose that $[\alpha]\psi \in \Gamma$. Since Γ is a TDS_n-MCS we know that $\langle\alpha\rangle\neg\psi \notin \Gamma$. By contraposition on our initial assumption, we have $\neg\psi \notin \Delta$. Because Δ is a TDS_n-MCS, we know that $\psi \in \Delta$. Since ψ was arbitrary, by the definition of $\mathcal{R}^c_{[\alpha]}$ (Definition 2.9) we have established that $\Delta \in \mathcal{R}^c_{[\alpha]}(\Gamma)$. Since both $\Delta, \Gamma \in \mathsf{IRR}$, by Definition 2.11 we know that $(\Gamma, \Delta) \in \mathcal{R}^{\mathsf{IRR}}_{[\alpha]} = \mathcal{R}^c_{[\alpha]} \cap \mathsf{IRR} \times \mathsf{IRR}$.

QED

It remains to show that the model $\mathfrak{M}^{\mathsf{IRR}}$ is, in fact, a TDS_n-model. It suffices to show that $\mathfrak{M}^{\mathsf{IRR}}$ satisfies the properties **C1-C6**, **D1-D4**, and **T1-T6** of Definition 2.4.

Lemma 2.11 (Canonical TDS_n-model)**.** *The canonical submodel $\mathfrak{M}^{\mathsf{IRR}}$ is a TDS_n-model.*

Proof. We must show that $\mathfrak{M}^{\mathsf{IRR}}$ satisfies properties **C1-C6**, **D1-D4**, and **T1-T6**. Due to the modularity of our approach, the proofs of the temporal properties **T1-T5**, as well as **C4-C6**, are those provided by Lorini (2013).[9] The proofs of **C1-C4**, **D1**, **D3**, and **D4** are straightforward adaptations of the ones given for the logic DS_n in Lemma 2.6. Thus, we only need to prove **D2**. In order to clarify how irreflexivity of $\mathfrak{M}^{\mathsf{IRR}}$ is guaranteed, we recall the proof of **T6** provided by Lorini (2013). Take an arbitrary $\Gamma \in \mathsf{IRR}$ of $\mathfrak{M}^{\mathsf{IRR}}$:

D2 We show that there exists a $\Delta \in \mathsf{IRR}$ such that $\Delta \in \mathcal{R}^{\mathsf{IRR}}_{\Box}(\Gamma)$ and for every $\Sigma \in \mathsf{IRR}$, if $\Sigma \in \mathcal{R}^{\mathsf{IRR}}_{[i]}(\Delta)$, then $\Sigma \in \mathcal{R}^{\mathsf{IRR}}_{\otimes_i}(\Gamma)$. Since Γ is an IRR-theory, there is a propositional variable p such that $name(p) \in \Gamma$. Define

$$\Delta_0 = \{[i]\varphi \mid \ \otimes_i \varphi \in \Gamma\} \cup \{\psi \mid \Box\psi \in \Gamma\} \cup \{name(p)\}$$

We prove by contradiction that Δ_0 is consistent and then extend Δ_0 to an IRR-theory. If Δ_0 is inconsistent, then

$$\vdash_{\mathsf{TDS}_n} ([i]\varphi_1 \wedge ... \wedge [i]\varphi_k \wedge \psi_1 \wedge ... \wedge \psi_l \wedge name(p)) \to \bot$$

where $[i]\varphi_1, \ldots, [i]\varphi_k \in \{[i]\varphi \mid \ \otimes_i \varphi \in \Gamma\}$ and $\psi_1, \ldots, \psi_l \in \{\psi \mid \Box\psi \in \Gamma\}$. Let $\hat{\varphi} = \varphi_1 \wedge ... \wedge \varphi_k$ and $\hat{\psi} = \psi_1 \wedge ... \wedge \psi_l$. Since, $\vdash_{\mathsf{TDS}_n} [i]\hat{\varphi} \equiv [i]\varphi_1 \wedge ... \wedge [i]\varphi_k$ we have

$$\vdash_{\mathsf{TDS}_n} (\hat{\psi} \wedge name(p)) \to \neg[i]\hat{\varphi}$$

By the normality of $\Box$ we know that $\vdash_{\mathsf{TDS}_n} \Box(\hat{\psi} \wedge name(p)) \to \Box\neg[i]\hat{\varphi}$, which implies $\vdash_{\mathsf{TDS}_n} \Box\hat{\psi} \wedge \Box name(p) \to \neg\Diamond[i]\hat{\varphi}$. Clearly, because $\Box\hat{\psi} \in \Gamma$, $name(p) \in \Gamma$ and $\vdash_{\mathsf{TDS}_n} name(p) \to \Box name(p)$, we have that $\Gamma \vdash_{\mathsf{TDS}_n} \neg\Diamond[i]\hat{\varphi}$. This implies that $\neg\Diamond[i]\hat{\varphi} \in \Gamma$ since Γ is an IRR-theory.

Also, since $\otimes_i\varphi_1, ..., \otimes_i\varphi_k \in \Gamma$ we have $\otimes_i\varphi_1 \wedge ... \wedge \otimes_i\varphi_k \in \Gamma$. By $\vdash_{\mathsf{TDS}_n} \otimes_i\hat{\varphi} \equiv \otimes_i\varphi_1 \wedge ... \wedge \otimes_i\varphi_k$ we conclude $\otimes_i\hat{\varphi} \in \Gamma$ as well. Since $\otimes_i\hat{\varphi} \to \Diamond[i]\hat{\varphi} \in \Gamma$ (axiom $\mathsf{A11}$), we obtain by modus ponens that $\Diamond[i]\hat{\varphi} \in \Gamma$. Since Γ is an IRR-theory (and hence consistent), we obtain a contradiction, which proves that Δ_0 is consistent.

We now extend Δ_0 to an IRR-theory Δ by first defining an increasing sequence Δ_0, Δ_1, ..., Δ_n, ... of sets of formulae. Suppose that Δ_n is consistent and defined, and enumerate the formulae of $\mathcal{L}^{td}_n$ so that each formula in odd position is an element of Zig. We define Δ_{n+1}.

Consider the formula ψ_n. Either, $\Delta_n \cup \{\neg\psi_n\}$ is consistent or $\Delta_n \cup \{\psi_n\}$ is consistent. If the former holds, then let $\Delta_{n+1} := \Delta_n \cup \{\neg\psi_n\}$. If the latter holds, then there are

[9] Although the non-deontic frame properties are the same, Lorini (2013) uses a different labeling of the properties than we do in this chapter. He defines the properties in Def. 2.4 of (Lorini, 2013). To facilitate comparison, we point out that **C1**, **C2**, and **C4** correspond to the second bullet in Def. 2.4. Properties **C3**, **C4**, and **C6** corresponds to (C1), (C2), respectively (C3) in Def. 2.4. The temporal properties **T1** and **T2** correspond to the third bullet in Def. 2.4. Last, **T3**, **T4**, **T5**, and **T6**, correspond to (C4), (C5), (C6), respectively (C7) in Def. 2.4.

two subcases to consider: either n is even, in which case, we let $\Delta_{n+1} := \Delta_n \cup \{\psi_n\}$, or n is odd, in which case, $\Delta_n \cup \{\psi_n\}$ is consistent and $\psi_n \in \mathsf{Zig}$. We show that in the latter subcase, we can find a propositional variable q such that $\Delta_n \cup \{\psi_n, \psi_n(q)\}$ is consistent; we then define $\Delta_{n+1} := \Delta_n \cup \{\psi_n, \psi_n(q)\}$.

First, we show that for Γ,

$$\ominus_i(\mathit{name}(p) \wedge \bigwedge_{\chi \in \Delta_n \setminus \Delta_0} \chi \wedge \psi_n) \in \Gamma \tag{2.1}$$

Suppose towards a contradiction that (2.1) does not hold. Then,

$$\otimes_i((\mathit{name}(p) \wedge \bigwedge_{\chi \in \Delta_n \setminus \Delta_0} \chi) \to \neg\psi_n) \in \Gamma$$

since Γ is an IRR-theory and has the properties specified by Lemma 2.2. By the definition of Δ_0 it follows that

$$[i]((\mathit{name}(p) \wedge \bigwedge_{\chi \in \Delta_n \setminus \Delta_0} \chi) \to \neg\psi_n) \in \Delta_n$$

Using the axiom scheme $[i]\theta \to \theta$ $\mathsf{A5}$, we infer that

$$\Delta_n \vdash_{\mathsf{TDS}_n} (\mathit{name}(p) \wedge \bigwedge_{\chi \in \Delta_n \setminus \Delta_0} \chi) \to \neg\psi_n$$

Since

$$\Delta_n \vdash_{\mathsf{TDS}_n} \mathit{name}(p) \wedge \bigwedge_{\chi \in \Delta_n \setminus \Delta_0} \chi$$

we conclude that $\Delta_n \vdash_{\mathsf{TDS}_n} \neg\psi_n$, which contradicts the fact that $\Delta_n \cup \{\psi_n\}$ is consistent and, so, (2.1) holds. Consequently, since Γ is an IRR-theory, we know that

$$\ominus_i(\mathit{name}(p) \wedge \bigwedge_{\chi \in \Delta_n \setminus \Delta_0} \chi \wedge \psi_n(q)) \in \Gamma \tag{2.2}$$

Using this fact, we prove that $\Delta_{n+1} := \Delta_n \cup \{\psi_n, \psi_n(q)\}$ is consistent. Suppose towards a contradiction otherwise. Then, there exist $\theta_1, \ldots, \theta_m \in \{\theta \mid \Box\theta \in \Gamma\}$ and $[i]\gamma_1, \ldots, [i]\gamma_k \in \{[i]\gamma \mid \otimes_i \gamma \in \Gamma\}$ such that

$$\vdash_{\mathsf{TDS}_n} \theta_1 \wedge \cdots \wedge \theta_m \to ([i]\gamma_1 \wedge \cdots \wedge [i]\gamma_k \to \neg(\mathit{name}(p) \wedge \bigwedge_{\chi \in \Delta_n \setminus \Delta_0} \chi \wedge \psi_n(q)))$$

By the normality of $\otimes_i$, we can derive

$$\vdash_{\mathsf{TDS}_n} \otimes_i(\theta_1 \wedge \cdots \wedge \theta_m) \to \otimes_i([i]\gamma_1 \wedge \cdots \wedge [i]\gamma_k \to \neg(\mathit{name}(p) \wedge \bigwedge_{\chi \in \Delta_n \setminus \Delta_0} \chi \wedge \psi_n(q)))$$

Using axiom $\mathsf{A12}$ we obtain

$$\vdash_{\mathsf{TDS}_n} \Box(\theta_1 \wedge \cdots \wedge \theta_m) \to \otimes_i([i]\gamma_1 \wedge \cdots \wedge [i]\gamma_k \to \neg(\mathit{name}(p) \wedge \bigwedge_{\chi \in \Delta_n \setminus \Delta_0} \chi \wedge \psi_n(q)))$$

By the assumption that $\theta_1, \ldots, \theta_m \in \{\theta \mid \Box\theta \in \Gamma\}$ and the fact that Γ is an IRR-theory, we know that $\Box(\theta_1 \wedge \cdots \wedge \theta_m) \in \Gamma$, implying that

$$\otimes_i([i]\gamma_1 \wedge \cdots \wedge [i]\gamma_k \rightarrow \neg(name(p) \wedge \bigwedge_{\chi \in \Delta_n \setminus \Delta_0} \chi \wedge \psi_n(q))) \in \Gamma$$

We infer by the normality of $\otimes_i$ that

$$\otimes_i[i](\gamma_1 \wedge \cdots \wedge \gamma_k) \rightarrow \otimes_i\neg(name(p) \wedge \bigwedge_{\chi \in \Delta_n \setminus \Delta_0} \chi \wedge \psi_n(q))) \in \Gamma$$

Using the axiom scheme $\otimes_i\varphi \rightarrow \otimes_i[i]\varphi$ (A13) we derive

$$\otimes_i(\gamma_1 \wedge \cdots \wedge \gamma_k) \rightarrow \otimes_i\neg(name(p) \wedge \bigwedge_{\chi \in \Delta_n \setminus \Delta_0} \chi \wedge \psi_n(q))) \in \Gamma$$

Our assumption implies that $\otimes_i(\gamma_1 \wedge \cdots \wedge \gamma_k) \in \Gamma$, and so

$$\otimes_i\neg(name(p) \wedge \bigwedge_{\chi \in \Delta_n \setminus \Delta_0} \chi \wedge \psi_n(q))) \in \Gamma$$

This contradicts (2.2) and proves that $\Delta_n \cup \{\psi_n\psi_n(q)\}$ is consistent.

It is straightforward to infer that Δ is an IRR-theory by an argument similar to Lemma 2.8.

Clearly, $\Delta \in \mathcal{R}_{\Box}^{\mathsf{IRR}}(\Gamma)$ holds by the definition of Δ. Last, let Σ be an arbitrary IRR-theory in IRR. Assume that $\Sigma \in \mathcal{R}_{[i]}^{\mathsf{IRR}}(\Delta)$ holds and let $\otimes_i\varphi \in \Gamma$. By definition $[i]\varphi \in \Delta$, and so, $\varphi \in \Sigma$ by the definition of the relation $\mathcal{R}_{[i]}^{\mathsf{IRR}}$, which completes the proof.

T6 Let $\Delta \in \mathsf{IRR}$ and assume that $\Delta \in \mathcal{R}_{\Box}^{\mathsf{IRR}}(\Gamma)$. We show that $\Delta \notin \mathcal{R}_{\mathsf{G}}^{\mathsf{IRR}}(\Gamma)$. Since Δ is an IRR-theory we know $name(p) \in \Delta$ for some atom $p \in \mathsf{Atoms}$. Consequently, $\Box\neg p, \Box\mathsf{G}p, \Box\mathsf{H}p \in \Delta$. Since $\Box\neg p \rightarrow \neg p \in \Delta$ (axiom A2) we also have $\neg p \in \Delta$. By the fact that $\mathcal{R}_{\Box}^{\mathsf{IRR}}$ is an equivalence relation, we have $\Gamma \in \mathcal{R}_{\Box}^{\mathsf{IRR}}(\Delta)$ and so $\mathsf{G}p \in \Gamma$. Furthermore, since Γ is an IRR-theory, we know that $\mathsf{F}\neg p \notin \Gamma$. Last, since $\neg p \in \Delta$ by Lemma 2.10 we obtain $\Delta \notin \mathcal{R}_{\mathsf{G}}^{\mathsf{IRR}}(\Gamma)$. QED

Theorem 2.3 (Weak completeness of TDS_n)**.** *For any formula $\varphi \in \mathcal{L}_n^{td}$, if $\models_{\mathsf{TDS}_n} \varphi$, then $\vdash_{\mathsf{TDS}_n} \varphi$.*

Proof. Suppose that $\varphi \in \mathcal{L}_n^{td}$ is consistent. By Lemma 2.8, we can extend φ to an IRR-theory Γ such that $\varphi \in \Gamma$. By Lemma 2.9, we know that the set IRR is a diamond saturated set, and so, by Lemma 2.7, we know that $\mathfrak{M}^{\mathsf{IRR}}, \Gamma \models \varphi$ iff $\varphi \in \Gamma$. Hence, we can conclude that $\mathfrak{M}^{\mathsf{IRR}}, \Gamma \models \varphi$. By Lemma 2.11 we know that $\mathfrak{M}^{\mathsf{IRR}}$ is a TDS_n-model. Therefore, φ is satisfiable on a TDS_n-model. QED

2.3 Transformations into Utilitarian Models

In this section, we prove that the logics TDS_n and DS_n are also sound and complete with respect to the traditional utilitarian semantics (Objective 3). This means that our relational semantics is equivalent to the utilitarian approach. In particular, we show that DS_n is equivalent to the logic of *dominance ought* based on (act) utilitarian STIT models, as developed by Horty (2001, Ch.4). We obtain these results by demonstrating how TDS_n-models can be truth-preservingly transformed into *utilitarian* STIT models. To be precise, we introduce a utility function 'util' that maps natural numbers—i.e., utilities—to worlds in the model.[10] In contrast to (Horty, 2001; Murakami, 2005), we start with assigning utilities to individual worlds and only later modify our approach by assigning utilities to complete histories (i.e., where all worlds on a timeline have the same utility). There are two reasons for doing this. First, the atemporal language of $\mathcal{L}_n^d$ cannot distinguish between multi-moment models and single-moment models (Balbiani et al., 2008; Murakami, 2005) and thus utility assignments may be safely restricted to individual worlds in DS_n. Second, once we move to an explicit temporal setting, certain problems arise with respect to assigning utilities to complete histories. The latter is discussed in Section 2.4.

2.3.1 The Semantics of Dominance Ought

Horty (2001) defines *Dominance Act Utilitarianism* as "a form of act utilitarianism applicable in the presence of both indeterminism and uncertainty, and based on the dominance ordering among actions" (p.73). Formally, indeterminism and uncertainty refer to branching time, respectively, choice in the context of STIT (see page 26). The act utilitarian approach to STIT takes the evaluation of utilities as the ground for obligation: by comparing utilities, one can obtain an ideality ordering on the choices available to each agent. Horty (2001, Ch.3-4) provides an extensive argument for the adaptation of, what he calls, *the dominance ought*: in brief, what an agent ought to see to is defined in terms of the choices that are not strongly dominated by any other choice available to the agent, *irrespective* of the choices made by any of the other agents. One may thus think of this notion as an all-things-considered obligation. In the remainder, we make the above formally precise and define $\underline{\text{T}}$emporal $\underline{\text{U}}$tilitarian $\underline{\text{S}}$TIT logic (for short, TUS_n). We use US_n to denote the atemporal subsystem called $\underline{\text{U}}$tilitarian $\underline{\text{S}}$TIT logic (for short, US_n).

Remark 2.2. *It must be noted that Horty (2001) initially developed the semantics of dominance ought for Branching Time frames (BT) with Agential Choice (AC) functions. Fortunately, because BT+AC frames can be directly translated into relational semantics (Balbiani et al., 2008; Lorini, 2013)—such as the one employed in this chapter—the trans-*

[10] Horty (2001) takes the reals as default utilities. Although irreflexive and serial branching time frames are infinite, they are countable infinite: every node is reachable by a finite path, and there is at most countably infinite branching at each node. Consequently, it suffices to use the natural numbers $\mathbb{N}$ to assign a (possibly unique) number to each world in the model. Last, we point out that the idea of a utility is abstract, i.e., how those utilities came about is not taken into consideration.

formation of the semantics of dominance ought (Horty, 2001) into relational semantics is straightforward.

Definition 2.13 (Frames and Models for TUS_n and US_n)**.** *A relational* Temporal Utilitarian STIT frame *(for short,* TUS_n*-frame) is a tuple* $\mathfrak{F} = \langle W, \mathcal{R}_\Box, \{\mathcal{R}_{[i]} | i \in Ag\}, \mathsf{util}, \mathcal{R}_{[Ag]}, \mathcal{R}_\mathsf{G}, \mathcal{R}_\mathsf{H}\rangle$*, where* $\mathfrak{F}$ *satisfies conditions **C1**–**C6** and **T1**–**T6** of Definition 2.4 together with the following:*

U1 $\mathsf{util} : W \mapsto \mathbb{N}$ *is a utility function assigning each world to a natural number.*

A TUS_n-model *is a tuple* $\mathfrak{M} = \langle \mathfrak{F}, V \rangle$ *where* $\mathfrak{F}$ *is a* TUS_n*-frame and* V *is a valuation function assigning propositional atoms to subsets of* W*, i.e.,* $V \colon \mathsf{Atoms} \mapsto \wp(W)$*.*

A relational Utilitarian STIT frame *(for short,* US_n*-frame) is a tuple* $\mathfrak{F} = \langle W, \mathcal{R}_\Box, \{\mathcal{R}_{[i]} | i \in Ag\}, \mathsf{util}\rangle$ *satisfying the conditions **C1-C4** of Definition 2.4 together with **U1** above. A* US_n-model *is defined as usual.*

In order to semantically characterize the interpretation of the dominance ought $\otimes_i$, we need some additional machinery. First of all, we need to make precise what it means for a choice to be optimal "*irrespective* of the choices made by any of the other agents". To model this, Horty introduces the notion of a *state* (of nature): "we will identify the states confronting an agent at any given moment with the possible patterns of action that might be performed at that moment by all other agents" (p.66). The principle of independence of agents ensures that no agent can influence the choices of any other agent. Therefore, one can regard the joint interaction of all other agents as a state of nature for that agent. Subsequently, an agent may compare each choice available to her with a given state, each resulting in a unique outcome (namely, that of all agents acting together). A dominance ordering then orders an agent's choices according to these possible outcomes.

Formally, let $v \in \mathcal{R}_\Box(w)$, then a *state* $\mathcal{R}^s_{[i]}(v)$ for agent i at v is defined as,

$$\mathcal{R}^s_{[i]}(v) \quad := \bigcap_{k \in \mathsf{Agents} \setminus \{i\}} \mathcal{R}_k(v)$$

The possible combinations of choices available to the set $\mathsf{Agents}\setminus\{i\}$ are the different states available at that moment to agent i. We point out that the independence of agents principle ensures that the joint choice of any combination of choices of all agents is non-empty, which, a fortiori, makes each individual choice, as well as each state, non-empty.

Subsequently, we define a *preference order* $\leq$ over choices and collective choices, including states. Let util be a function assigning natural numbers to worlds, i.e., $\mathsf{util} : W \mapsto \mathbb{N}$, and let $\mathcal{R}_{[i]}(v), \mathcal{R}_{[i]}(z) \subseteq \mathcal{R}_\Box(w)$, then weak preference is defined as[11],

[11] Henceforth, we use the quantifies ∀ and ∃ as abbreviations for 'for all' and 'there exists a'.

$$\mathcal{R}_{[i]}(v) \leq \mathcal{R}_{[i]}(z) := \forall v^* \in \mathcal{R}_{[i]}(v), \forall z^* \in \mathcal{R}_{[i]}(z), \mathsf{util}(v^*) \leq \mathsf{util}(z^*)$$

where $\mathsf{util}(v)$ denotes the natural number assigned to v. The preference order over $\mathcal{R}^s_{[i]}$ is defined similarly. The above definition states that, for an agent i, a choice is weakly preferred over another whenever all values of the possible outcomes of the former are at least as high as those of the latter. Strict preference is then defined as usual, i.e., $\mathcal{R}_{[i]}(v) < \mathcal{R}_{[i]}(z) := \mathcal{R}_{[i]}(v) \leq \mathcal{R}_{[i]}(z)$ and $\mathcal{R}_{[i]}(z) \not\leq \mathcal{R}_{[i]}(v)$.

Next, a *dominance order* $\preceq$ over choices $\mathcal{R}_{[i]}(v), \mathcal{R}_{[i]}(z) \subseteq \mathcal{R}_{\Box}(w)$ is defined as,

$$\mathcal{R}_{[i]}(v) \preceq \mathcal{R}_{[i]}(z) := \forall \mathcal{R}^s_{[i]}(x) \subseteq \mathcal{R}_{\Box}(w), \mathcal{R}_{[i]}(v) \cap \mathcal{R}^s_{[i]}(x) \leq \mathcal{R}_{[i]}(z) \cap \mathcal{R}^s_{[i]}(x)$$

where $\forall \mathcal{R}^s_{[i]}(x) \subseteq \mathcal{R}_{\Box}(w)$ means for each available state to i at moment $\mathcal{R}_{\Box}(w)$ with $x \in \mathcal{R}_{\Box}(w)$. Informally, the dominance ordering expresses that an agent's choice weakly dominates another if the values of the outcomes of the former are weakly preferred to those of the latter choice, *given any possible state available to that agent.* Last, *strict dominance* is defined as usual, i.e., $\mathcal{R}_{[i]}(v) \prec \mathcal{R}_{[i]}(z) := \mathcal{R}_{[i]}(v) \preceq \mathcal{R}_{[i]}(z)$ and $\mathcal{R}_{[i]}(z) \not\preceq \mathcal{R}_{[i]}(v)$. We use the dominance ordering for the semantic evaluation of the modal operator $\otimes_i$ (see (Horty, 2001, Ch.4) for a more detailed discussion).

Definition 2.14 (Semantics of TUS_n- and US_n-models)**.** *Let $\mathfrak{M}$ be a TUS_n-model, $w \in W$ of $\mathfrak{M}$ and let $\|\varphi\| = \{w \mid \mathfrak{M}, w \models \varphi\}$ be the* truth-set *of φ over $\mathfrak{M}$. We define* satisfaction *of a formula $\varphi \in \mathcal{L}^{td}_n$ at a world w of $\mathfrak{M}$ by adopting clauses 1–5 and 7–9 of Definition 2.5 together with the following clause:*

10. $\mathfrak{M}, w \models \otimes_i\varphi$ iff for all $\mathcal{R}_{[i]}(v) \subseteq \mathcal{R}_{\Box}(w)$, if $\mathcal{R}_{[i]}(v) \not\subseteq \|\varphi\|$, then there is a $\mathcal{R}_{[i]}(z) \subseteq \mathcal{R}_{\Box}(w)$ such that (i) $\mathcal{R}_{[i]}(v) \prec \mathcal{R}_{[i]}(z)$, (ii) $\mathcal{R}_{[i]}(z) \subseteq \|\varphi\|$, and (iii) for each $\mathcal{R}_{[i]}(x) \subseteq \mathcal{R}_{\Box}(w)$, if $\mathcal{R}_{[i]}(z) \preceq \mathcal{R}_{[i]}(x)$, then $\mathcal{R}_{[i]}(x) \subseteq \|\varphi\|$.

Let $\mathfrak{M}$ be a US_n-model. We define satisfaction *of a formula $\varphi \in \mathcal{L}^d_s$ by adopting clauses 1–5 of Definition 2.5 together with clause 10 above.*

Global truth, frame validity, and semantic entailment are defined as usual (Definition 2.4). We define the logic TUS_n *as the set of all $\mathcal{L}^{td}_n$ formulae valid on the class of TUS_n-models. The* logic US_n *is defined as the set of all $\mathcal{L}^d_n$ formulae valid on the class of US_n-models.*

Clause 10 of Definition 2.14 is a relational representation of the semantic evaluation of the *dominance ought* in (Horty, 2001). It must be interpreted as follows: agent i ought to see to it that φ holds iff for every choice $\mathcal{R}_{[i]}(v)$ available to i that does not guarantee φ there (i) exists a strictly dominating choice $\mathcal{R}_{[i]}(z)$ that (ii) does guarantee φ and (iii) every weakly dominating choice $\mathcal{R}_{[i]}(x)$ over $\mathcal{R}_{[i]}(z)$ also guarantees φ. In other words, all choices not guaranteeing φ are strictly dominated by choices guaranteeing φ at the moment of evaluation.

Notice that TUS_n-frames only differ from TDS_n-frames through replacing the relation $\mathcal{R}_{\otimes_i}$ (for each $i \in \mathsf{Agents}$) and corresponding conditions **D1**–**D4** with the utility function util and condition **U1**. The same holds true for US_n- and DS_n-frames.

Example 2.3 (A Utilitarian Scenario)**.** *Consider the utility assignment in Figure 2.2 on page 36. The utilities at moment $\mathcal{R}_\Box(v_i)$ are assigned as follows: $\mathsf{util}(v_i) = 4$ for $i \in \{1,2,3\}$ and $\mathsf{util}(v_4) = 3$.*[12] *At $\mathcal{R}_\Box(v_i)$, both John and Paul have the choice to thank each other for working it out, i.e., $\Diamond[j]$`thank_j` and $\Diamond[p]$`thank_p`. In fact, for both agents, this choice guarantees a utility of 4 and John and Paul are obliged to thank one another. To illustrate, using Definition 2.14, we know that John ought to see to it that he thanks Paul—i.e., $\otimes_j$`thank_j`—since the choice $\mathcal{R}_{[j]}(v_2)$ not guaranteeing `thank_j` is strictly dominated by the only other choice $\mathcal{R}_{[j]}(v_1)$ guaranteeing `thank_j` (to see this, observe that $v_4 \in \mathcal{R}_{[j]}(v_2)$ has a utility of 3). If the two agents fulfill their duty, this yields a utility of 4 at $\mathcal{R}_\Box(v_i)$. If both act against their duty, the outcome will be of a strictly lesser utility 3.*

2.3.2 Equivalence of the Two Semantics

First, we show that the Hilbert-style axiomatizations TDS_n and DS_n are sound with respect to the class of TUS_n-models, respectively, the class of US_n-models. We start by pointing out some useful facts.

Lemma 2.12. *Let DS_n^- be the atemporal* minimal *deontic* STIT *logic consisting of axioms* $\mathsf{A0}$-$\mathsf{A9}$ *and* $\mathsf{A12}$*, and the rules* $\mathsf{R0}$ *and* $\mathsf{R1}$*. The following holds:*

1. *Let $\mathsf{DS}_n^-\{\mathsf{A10}\}$ be the logic DS_n^- extended with axiom $\mathsf{A10}$ $\otimes_i\varphi \to \Box\otimes_i\varphi$ and let $\mathsf{DS}_n^-\{\mathsf{B10}\}$ be the logic DS_n^- extended with axiom $\mathsf{B10}$ $\Diamond\otimes_i\varphi \to \Box\otimes_i\varphi$. Then: $\mathsf{DS}_n^-\{\mathsf{A10}\} \equiv \mathsf{DS}_n^-\{\mathsf{B10}\}$.*

2. *Let $\mathsf{DS}_n^-\{\mathsf{A13}\}$ be the logic DS_n^- extended with axiom $\mathsf{A13}$ $\otimes_i\varphi \to \otimes_i[i]\varphi$ and let $\mathsf{DS}_n^-\{\mathsf{B13}\}$ be the logic DS_n^- extended with axiom $\mathsf{B13}$ $\Box([i]\varphi \to [i]\psi) \to (\otimes_i\varphi \to \otimes_i\psi)$. Then: $\mathsf{DS}_n^-\{\mathsf{A13}\} \equiv \mathsf{DS}_n^-\{\mathsf{B13}\}$.*

Proof. The proofs are straightforward cases of modal reasoning. We briefly sketch the main steps and theorems used.

Ad (1). The left-to-right direction straightforwardly follows from the fact that $\Box$ is an $\mathsf{S5}$ modality in DS_n^- and, so, $\otimes_i\varphi \to \Diamond\otimes_i\varphi$ is a $\mathsf{DS}_n^-\{\mathsf{B10}\}$-theorem. For the right-to-left direction, observe that $\Diamond\otimes_i\varphi \to \Diamond\Box\otimes_i\varphi$ is a $\mathsf{DS}_n^-\{\mathsf{A10}\}$-theorem by the normality of $\Box$. By the fact that $\Box$ is an $\mathsf{S5}$ operator, we know $\Diamond\Box\otimes_i\varphi \to \Box\otimes_i\varphi$ is a $\mathsf{DS}_n^-\{\mathsf{A10}\}$-theorem and so, by basic modal reasoning, $\Diamond\otimes_i \to \Box\otimes_i\varphi$ is a theorem.

[12] In Figure 2.2 utilities are represented as assigned to histories. In Section 2.4, we reconsider the example in light of utility functions restricted to histories. We note that each utility function restricted to histories straightforwardly yields a function restricted to moments. We refer to Definition 2.16 for details.

Ad (2). The left-to-right direction follows from the basic STIT theorem $\Box([i]\varphi \to [i][i]\varphi)$ (S5 behavior of $[i]$), together with $\Box([i]\varphi \to [i]\psi) \to (\otimes_i\varphi \to \otimes_i\psi)$, which implies $\otimes_i\varphi \to \otimes_i[i]\varphi$ as a $\mathsf{DS}_n^-\{\mathsf{B13}\}$-theorem. For the right-to-left direction, it suffices the observe the following: $\Box([i]\varphi \to [i]\psi), \otimes_i\varphi \vdash_{\mathsf{DS}_n^-\{\mathsf{A13}\}} \otimes_i([i]\varphi \to [i]\psi) \land \otimes_i[i]\varphi$ (by A12 and A13, respectively). The normality of $\otimes_i$ implies $\Box([i]\varphi \to [i]\psi), \otimes_i\varphi \vdash_{\mathsf{DS}_n^-\{\mathsf{A13}\}} \otimes_i(([i]\varphi \to [i]\psi) \land [i]\varphi)$. By straightforward modal reasoning we have $\Box([i]\varphi \to [i]\psi), \otimes_i\varphi \vdash_{\mathsf{DS}_n^-\{\mathsf{A13}\}} \otimes_i[i]\psi$. By the fact that $[i]$ is an S5 modality and by the normality of $\otimes_i$ we know $\vdash_{\mathsf{DS}_n^-\{\mathsf{A13}\}} \otimes_i[i]\psi \to \otimes_i\psi$. Consequently, $\Box([i]\varphi \to [i]\psi), \otimes_i\varphi \vdash_{\mathsf{DS}_n^-\{\mathsf{A13}\}} \otimes_i\psi$. QED

Lemma 2.12 demonstrates that the alternative axiomatization of the deontic STIT modality $\otimes_i$ by Murakami (2005) is equivalent to the axiomatization of $\otimes_i$ provided in this chapter (Definition 2.2). We use this fact in the following theorem.

Theorem 2.4 (Soundness of TUS_n)**.** *For each $\varphi \in \mathcal{L}_n^{td}$ and $\Gamma \subseteq \mathcal{L}_n^{td}$, if $\Gamma \vdash_{\mathsf{TDS}_n} \varphi$, then $\Gamma \models_{\mathsf{TUS}_n} \varphi$.*

Proof. By the modularity of our approach, it suffices to only consider the deontic axioms A10-A13. The axioms A11 and A12 were shown sound by Murakami (2005) with respect to US_n and hence they are also sound with respect to TUS_n. Furthermore, the axioms $\Diamond\otimes_i \to \Box \otimes_i \varphi$ and $\Box([i]\varphi \to [i]\psi) \to (\otimes_i\varphi \to \otimes_i\psi)$ are shown sound by (Murakami, 2005) and so, by Lemma 2.12, we know that axioms A10 and A13 are sound with respect to TUS_n too. QED

Completeness is shown through the stronger result in Theorem 2.6, which demonstrates that the class of TUS_n-models characterizes the same set of formulae as the class of TDS_n-models. In what follows, we make (often implicit) use of the following lemma.

Lemma 2.13. *The following holds for any TUS_n-, respectively TDS_n-model:*

1. *For all $v \in \mathcal{R}^s_{[i]}(w)$, we have $\mathcal{R}^s_{[i]}(w) = \mathcal{R}^s_{[i]}(v)$;*
2. *For all $\mathcal{R}_{[i]}(z) \subseteq \mathcal{R}_\Box(w)$, either $\mathcal{R}_{[i]}(z) \subseteq \mathcal{R}_{\otimes_i}(w)$ or $\mathcal{R}_{[i]}(z) \cap \mathcal{R}_{\otimes_i}(w) = \emptyset$.*

Proof. Claim (1) follows from the fact that $\mathcal{R}^s_{[i]}$ is an equivalence class. We prove (2) by reasoning towards a contradiction. Suppose there is a $\mathcal{R}_{[i]}(z) \subseteq \mathcal{R}_\Box(w)$ such that $\mathcal{R}_{[i]}(z) \not\subseteq \mathcal{R}_{\otimes_i}(w)$ and $\mathcal{R}_{[i]}(z) \cap \mathcal{R}_{\otimes_i}(w) \neq \emptyset$. It follows that there is a $z' \in \mathcal{R}_{[i]}(z) \setminus \mathcal{R}_{\otimes_i}(w)$ and, consequently, $z' \notin \mathcal{R}_{\otimes_i}(w)$. Furthermore, there is a $z'' \in \mathcal{R}_{[i]}(z) \cap \mathcal{R}_{\otimes_i}(w)$ which implies that $z'' \in \mathcal{R}_{\otimes_i}(w)$. By Lemma 2.1-(2) we know that $z' \in \mathcal{R}_{[i]}(z) = \mathcal{R}_{[i]}(z'')$ and, consequently, by **D4** we obtain $z' \in \mathcal{R}_{\otimes_i}(w)$. Contradiction. QED

We now prove that every TUS_n-valid formula is a TDS_n-valid formula (Lemma 2.16). We do this by constructing a TUS_n model from a TDS_n-model (Lemma 2.14). Then, we show that the constructed TUS_n-model satisfies exactly the same formulae as the TDS_n-model from which it is obtained (Lemma 2.15). We start by defining the transformation.

Definition 2.15. *Let* $\mathcal{M}^{td} = \langle \mathcal{W}, \mathcal{R}_{\Box}, \{\mathcal{R}_{[i]} \mid i \in \mathsf{Agents}\}, \{\mathcal{R}_{\otimes_i} | i \in \mathsf{Agents}\}, \mathcal{R}_{[Ag]}, \mathcal{R}_{\mathsf{G}}, \mathcal{R}_{\mathsf{H}}, \mathcal{V}\rangle$ *be a* TDS_n*-model. Let* $\mathsf{M}^u = \langle \mathsf{W}, \mathsf{R}_{\Box}, \{\mathsf{R}_{[i]} | i \in \mathsf{Agents}\}, \mathsf{util}, \mathsf{R}_{[Ag]}, \mathsf{R}_{\mathsf{G}}, \mathsf{R}_{\mathsf{H}}, \mathsf{V}\rangle$ *be defined as follows:* $\mathsf{W} := \mathcal{W}$*,* $\mathsf{R}_{\Box} := \mathcal{R}_{\Box}$*,* $\mathsf{R}_{[i]} := \mathcal{R}_{[i]}$*,* $\mathsf{R}_{[Ag]} := \mathcal{R}_{[Ag]}, \mathsf{R}_{\mathsf{G}} := \mathcal{R}_{\mathsf{G}}, \mathsf{R}_{\mathsf{H}} := \mathcal{R}_{\mathsf{H}}$*, and* $\mathsf{V}(p) := \mathcal{V}(p)$ *for each* $p \in \mathsf{Atoms}$*. Let* util *be a function assigning each* $w \in \mathsf{W}$ *to a natural number* $i \in \mathbb{N}$ *according to the following three criteria:*

u1. For all $i \in \mathsf{Agents}$*, and for all* $w, v, z \in \mathcal{W}$*, if* $v, z \in \mathcal{R}_{\Box}(w)$*,* $v \in \mathcal{R}^s_{[i]}(w) \setminus \mathcal{R}_{\otimes_i}(w)$*, and* $z \in \mathcal{R}^s_{[i]}(w) \cap \mathcal{R}_{\otimes_i}(w)$*, then* $\mathsf{util}(v) \leq \mathsf{util}(z)$*;*

u2. For all $w, v, z \in \mathcal{W}$*, if* $v \in \mathcal{R}_{\Box}(w) \backslash \mathcal{R}_{\otimes_{Ag}}(w)$ *and* $z \in \mathcal{R}_{\otimes_{Ag}}(w)$*, then* $\mathsf{util}(v) < \mathsf{util}(z)$*;*

u3. For all $w, u, z \in W$*, if* $v, z \in \mathcal{R}^s_{[i]}(w) \cap \mathcal{R}_{\otimes_i}(w)$*, then* $\mathsf{util}(v) = \mathsf{util}(z)$*.*

where $\mathcal{R}_{\otimes_{Ag}} := \bigcap_{i \in \mathsf{Agents}} \mathcal{R}_{\otimes_i}$*.*

To enhance the readability of our proofs, we briefly discuss the intuition behind the three criteria. In the sequel, we call a world $w \in \mathcal{R}_{\otimes_i}$ a deontically ideal world. Then, u1 expresses that all deontically ideal worlds belonging to a particular state have a utility at least as high as any non-deontically ideal world belonging to that same state; u2 stipulates that those worlds deontically ideal for all the agents have a strictly higher utility than any other world; and u3 ensures that all deontically ideal worlds belonging to the same state receive the same utility.

The following lemma shows that the obtained model is, in fact, a TUS_n-model.

Lemma 2.14. *Let* $\mathcal{M}^{td}$ *be a* TDS_n*-model and let* M^u *be obtained following Definition 2.15:* M^u *is a* TUS_n*-model*

Proof. Observe that conditions **C1**–**C6**, and **T1**–**T6** of Definition 2.4 are satisfied by M^u since all of the relations of $\mathcal{M}^{td}$, with the exception of $\mathcal{R}_{\otimes_i}$, are identical to those in M^u. Furthermore, util satisfies property **U1** of Definition 2.13 and is well-defined. QED

Lemma 2.15 shows that satisfaction in the constructed TUS_n-model is equivalent to the TDS_n-model from which it is generated. Since the proof is for arbitrary TDS_n-models we know that a function util of Definition 2.15 exists for every such model.

Lemma 2.15. *Let* $\mathcal{M}^{td}$ *be a* TDS_n *model and let* M^u *be obtained following Definition 2.15. For all* $\psi \in \mathcal{L}^{td}_n$ *and all* $w \in \mathcal{W}$*:* $\mathcal{M}^{td}, w \models \psi$ iff $\mathsf{M}^u, w \models \psi$*.*

Proof. The proof is by induction on the complexity of ψ.

Base Case $\varphi = p$*.* By the definition of V in M^u it follows directly that $\mathcal{M}^{td}, w \models p$ iff $w \in \mathcal{V}$ iff $w \in \mathsf{V}$ iff $\mathsf{M}^u, w \models p$.

Inductive Step. The cases for the propositional connectives and the modalities $[\alpha] \in \{\Box\} \cup \{[i] \mid i \in \mathsf{Agents}\} \cup \{[Ag], \mathsf{G}, \mathsf{H}\}$ are straightforward by the construction of M^u. We consider the only non-trivial case $\psi = \otimes_i \varphi$.

Left-to-right. Assume $\mathcal{M}^{td}, w \models \otimes_i \varphi$. By the semantic interpretation of $\otimes_i$ in Definition 2.13 it suffices to prove that: for all $\mathsf{R}_{[i]}(v) \subseteq \mathsf{R}_\Box(w)$, if $\mathsf{R}_{[i]}(v) \not\subseteq \|\varphi\|_{\mathsf{M}^u}$, then there is a $\mathsf{R}_{[i]}(u) \subseteq \mathsf{R}_\Box(w)$ such that the following three clauses hold:

(i) $\mathsf{R}_{[i]}(v) \prec \mathsf{R}_{[i]}(u)$;

(ii) $\mathsf{R}_{[i]}(u) \subseteq \|\varphi\|_{\mathsf{M}^u}$;

(iii) for all $\mathsf{R}_{[i]}(x) \subseteq \mathsf{R}_\Box(w)$, if $\mathsf{R}_{[i]}(u) \preceq \mathsf{R}_{[i]}(x)$ then $\mathsf{R}_{[i]}(x) \subseteq \|\varphi\|_{\mathsf{M}^u}$.

Let $\mathsf{R}_{[i]}(v) \subseteq \mathsf{R}_\Box(w)$ be an arbitrary choice and assume that $\mathsf{R}_{[i]}(v) \not\subseteq \|\varphi\|_{\mathsf{M}^u}$. We prove the existence of a choice $\mathsf{R}_{[i]}(u) \subseteq \mathsf{R}_\Box(w)$ for which conditions (i)–(iii) hold. Observe that since $\mathcal{M}^{td}$ satisfies **C3** and **D2** of Definition 2.4 we know that

$$\text{there is a } u \in \mathcal{W} \text{ such that } \mathcal{R}_{[i]}(u) \subseteq \mathcal{R}_\Box(w) \text{ and } \mathcal{R}_{[i]}(u) \subseteq \mathcal{R}_{\otimes_i}(w) \tag{2.3}$$

and, therefore, by construction of M^u we know $\mathcal{R}_{[i]}(u) = \mathsf{R}_{[i]}(u)$. We demonstrate that conditions (i)–(iii) hold for $\mathsf{R}_{[i]}(u) \subseteq \mathsf{R}_\Box(w)$.

Before we address each item, we make two useful observations concerning $\mathcal{R}_{[i]}(u)$. By **D2** we know that for all $j \in \mathsf{Agents} \backslash \{i\}$, there is a $u_j \in \mathcal{R}_\Box(w)$ such that $\mathcal{R}_{[j]}(u_j) \subseteq \mathcal{R}_{\otimes_j}(w)$ and by **C4** (IoA) we have $\bigcap_{j \in \mathsf{Agents} \backslash \{i\}} \mathcal{R}_{[j]}(u_j) \cap \mathcal{R}_{[i]}(u) \neq \emptyset$. Therefore,

$$\text{there exists a } u^* \in \bigcap_{j \in \mathsf{Agents} \backslash \{i\}} \mathcal{R}_{[j]}(u_j) \cap \mathcal{R}_{[i]}(u). \tag{2.4}$$

As a consequence, the following statement holds for u^* at $\mathcal{M}^{td}$:

$$u^* \in \bigcap_{j \in \mathsf{Agents} \backslash \{i\}} \mathcal{R}_{\otimes_j}(w) \cap \mathcal{R}_{\otimes_i}(w) = \mathcal{R}_{\otimes_{Ag}}(w). \tag{2.5}$$

We now prove (i)–(iii):

(i) We show that $\mathsf{R}_{[i]}(v) \prec \mathsf{R}_{[i]}(u)$, that is, we show (a) $\mathsf{R}_{[i]}(v) \preceq \mathsf{R}_{[i]}(u)$ and (b) $\mathsf{R}_{[i]}(u) \not\preceq \mathsf{R}_{[i]}(v)$.

(a) Recall our assumption that $\mathsf{R}_{[i]}(v) \not\subseteq \|\varphi\|_{\mathsf{M}^u}$. Thus, we know there is a $v^* \in \mathsf{R}_{[i]}(v)$ s.t. $\mathsf{M}^u, v^* \not\models \varphi$. By construction of M^u, $v^* \in \mathcal{R}_{[i]}(v)$ and by IH we have $\mathcal{M}^{td}, v^* \not\models \varphi$. Consequently, by the assumption that $\mathcal{M}^{td}, w \models \otimes_i \varphi$, and the fact that $\mathcal{M}^{td}, v^* \not\models \varphi$, it follows that $v^* \notin \mathcal{R}_{\otimes_i}(w)$. Hence, we know that $\mathcal{R}_{[i]}(v) \not\subseteq \mathcal{R}_{\otimes_i}(w)$, which implies

$\mathcal{R}_{\otimes_i}(w) \cap \mathcal{R}_{[i]}(v) = \emptyset$ by Lemma 2.13–(2). Therefore, by $\mathcal{R}_{\otimes_i}(w) \cap \mathcal{R}_{[i]}(v) = \emptyset$ along with statement (2.3), we know that

$$\begin{aligned}&\text{For all } x, u', v' \in \mathcal{W}, \text{ if } v' \in \mathcal{R}^s_{[i]}(x) \cap \mathcal{R}_{[i]}(v) \text{ and } u' \in \mathcal{R}^s_{[i]}(x) \cap \mathcal{R}_{[i]}(u),\\ &\text{then } v' \in \mathcal{R}^s_{[i]}(x) \backslash \mathcal{R}_{\otimes_i}(w) \text{ and } u' \in \mathcal{R}^s_{[i]}(x) \cap \mathcal{R}_{\otimes_i}(w).\end{aligned} \tag{2.6}$$

Let $x, u', v' \in \mathcal{W}$ be arbitrary and assume that $v' \in \mathcal{R}^s_{[i]}(x) \cap \mathcal{R}_{[i]}(v)$ and $u' \in \mathcal{R}^s_{[i]}(x) \cap \mathcal{R}_{[i]}(u)$. By statement (2.6), it follows that $v' \in \mathcal{R}^s_{[i]}(x) \backslash \mathcal{R}_{\otimes_i}(w)$ and $u' \in \mathcal{R}^s_{[i]}(x) \cap \mathcal{R}_{\otimes_i}(w)$, which in conjunction with criterion u1 of the util function of M^u (Definition 2.15) implies that $\mathsf{util}(v') \leq \mathsf{util}(u')$. Therefore, the following holds,

$$\begin{aligned}&\text{For all } x, u', v' \in \mathcal{W}, \text{ if } v' \in \mathcal{R}^s_{[i]}(x) \cap \mathcal{R}_{[i]}(v) \text{ and } u' \in \mathcal{R}^s_{[i]}(x) \cap \mathcal{R}_{[i]}(u),\\ &\text{then } \mathsf{util}(v') \leq \mathsf{util}(u').\end{aligned} \tag{2.7}$$

It follows that for all $\mathcal{R}^s_{[i]}(x) \subseteq \mathcal{R}_{\Box}(w)$, $\mathcal{R}^s_{[i]}(x) \cap \mathcal{R}_{[i]}(v) \leq \mathcal{R}^s_{[i]}(x) \cap \mathcal{R}_{[i]}(u)$. Hence, by the definition of $\preceq$ and the definition of M^u, we obtain $\mathsf{R}_{[i]}(v) \preceq \mathsf{R}_{[i]}(u)$.

(b) We need to show $\mathsf{R}_{[i]}(u) \npreceq \mathsf{R}_{[i]}(v)$. By definition of $\preceq$, it suffices to show that there are $x, u', v' \in \mathsf{W}$ such that $\mathsf{R}_{[i]}(x) \subseteq \mathsf{R}_{\Box}(w)$, $u' \in \mathsf{R}_{[i]}(u) \cap \mathsf{R}^s_{[i]}(x)$, $v' \in \mathsf{R}_{[i]}(v) \cap \mathsf{R}^s_{[i]}(x)$, and $\mathsf{util}(v') < \mathsf{util}(u')$. Consider $\bigcap_{j \in \mathsf{Agents} \backslash \{i\}} \mathcal{R}_{[j]}(u_j) \cap \mathcal{R}_{[i]}(u) \neq \emptyset$ from statement (2.5). Let $\mathsf{R}^s_{[i]}(x) = \bigcap_{j \in \mathsf{Agents} \backslash \{i\}} \mathsf{R}_{[j]}(u_j) = \bigcap_{j \in \mathsf{Agents} \backslash \{i\}} \mathcal{R}_{[i]}(u_j) = \mathcal{R}^s_{[i]}(x)$. Clearly, $\mathsf{R}^s_{[i]}(x) \subseteq \mathsf{R}_{\Box}(w)$. Since $\mathcal{M}^{td}$ satisfies **C4** (IoA) we know that $\mathcal{R}^s_{[i]}(x) \cap \mathcal{R}_{[i]}(v) \neq \emptyset$ and so $\mathsf{R}^s_{[i]}(x) \cap \mathsf{R}_{[i]}(v) \neq \emptyset$ by the definition of M^u. Therefore, there is a $v' \in \mathsf{R}^s_{[i]}(x) \cap \mathsf{R}_{[i]}(v)$. Since $u^* \in \bigcap_{j \in \mathsf{Agents} \backslash \{i\}} \mathcal{R}_{[j]}(u_j) \cap \mathcal{R}_{[i]}(u)$ by statement (2.4), we know that $u^* \in \bigcap_{j \in \mathsf{Agents} \backslash \{i\}} \mathsf{R}_{[j]}(u_j) \cap \mathsf{R}_{[i]}(u)$, implying that $u^* \in \mathsf{R}^s_{[i]}(x) \cap \mathsf{R}_{[i]}(u)$. We know $\mathcal{R}_{\otimes_i}(w) \cap \mathcal{R}_{[i]}(v) = \emptyset$ by Lemma 2.13-(2) and thus $\mathcal{R}_{[i]}(v) \cap \mathcal{R}_{\otimes_{Ag}}(w) = \emptyset$ too. Consequently, since $\mathcal{R}_{[i]}(v) \neq \emptyset$ we know there is a $v' \in \mathcal{R}_{[i]} \cap \mathcal{R}_{\Box}(w) \setminus \mathcal{R}_{\otimes_{Ag}}(w)$. By criterion u2 of the util function of M^u (Definition 2.15) and the facts $v' \in \mathcal{R}_{\Box}(w) \setminus \mathcal{R}_{\otimes_{Ag}}(w)$ and $u^* \in \mathcal{R}_{\otimes_{Ag}}(w)$ (statement (2.5)), we have $\mathsf{util}(v') < \mathsf{util}(u^*)$. Therefore, $\mathsf{R}_{[i]}(u) \npreceq \mathsf{R}_{[i]}(v)$.

(ii) By assumption $\mathcal{R}_{\otimes_i}(w) \subseteq \|\varphi\|_{\mathcal{M}^{td}}$ and by statement (2.5) we know $\mathcal{R}_{[i]}(u) \subseteq \mathcal{R}_{\otimes_i}(w)$. By IH we have $\|\varphi\|_{\mathcal{M}^{td}} = \|\varphi\|_{\mathsf{M}^u}$ and since $\mathcal{R}_{[i]}(u) = \mathsf{R}_{[i]}(u)$ we know $\mathsf{R}_{[i]}(u) \subseteq \|\varphi\|_{\mathsf{M}^u}$.

(iii) We prove the case by contraposition and show that for all $\mathsf{R}_{[i]}(x) \subseteq \mathsf{R}_{\Box}(w)$, if $\mathsf{R}_{[i]}(x) \nsubseteq \|\varphi\|_{\mathsf{M}^u}$, then $\mathsf{R}_{[i]}(u) \npreceq \mathsf{R}_{[i]}(x)$. Let $\mathsf{R}_{[i]}(x)$ be an arbitrary choice in $\mathsf{R}_{\Box}(w)$ and assume that $\mathsf{R}_{[i]}(x) \nsubseteq \|\varphi\|_{\mathsf{M}^u}$. We prove that $\mathsf{R}_{[i]}(u) \npreceq \mathsf{R}_{[i]}(x)$. By definition of $\preceq$ it suffices to show that there is a state $\mathsf{R}^s_{[i]}(y) \subseteq \mathsf{R}_{\Box}(w)$ such that there is a $u' \in \mathsf{R}_{[i]}(u) \cap \mathsf{R}^s_{[i]}(y)$ and a $x' \in \mathsf{R}_{[i]}(x) \cap \mathsf{R}^s_{[i]}(y)$, with $\mathsf{util}(x') < \mathsf{util}(u')$. We prove that $\mathsf{R}^s_{[i]}(u^*)$ is this state.

By the assumption that $\mathsf{R}_{[i]}(x) \nsubseteq \|\varphi\|_{\mathsf{M}^u}$, we know there is a $x' \in \mathsf{R}_{[i]}(x)$ such that $\mathsf{M}^u, x' \not\models \varphi$. Clearly, $x' \in \mathcal{R}_{[i]}(x)$, and by IH we know that $\mathcal{M}^{td}, x' \not\models \varphi$. Since $\mathcal{M}^{td}, w \models \otimes_i \varphi$, we obtain $x' \notin \mathcal{R}_{\otimes_i}(w)$, and by Lemma 2.13–(2) we obtain

$$\mathcal{R}_{[i]}(x) \not\subseteq \mathcal{R}_{\otimes_i}(w). \tag{2.8}$$

By statement (2.5) we have $u^* \in \mathcal{R}_{\otimes_{Ag}}(w)$ and $u^* \in \mathcal{R}_{\otimes_i}(w)$. Also, we know $u^* \in \mathcal{R}_{[i]}(u)$ by statement (2.4). Since, $u^* \in \bigcap_{j \in Ag \setminus \{i\}} \mathcal{R}_{[j]}(u_j) \cap \mathcal{R}_{[i]}(u)$, we also have $u^* \in \bigcap_{j \in \mathsf{Agents} \setminus \{i\}} \mathcal{R}_{[j]}(u_j)$. Let $\mathcal{R}^s_{[i]}(u^*) = \bigcap_{j \in Ag \setminus \{i\}} \mathcal{R}_{[j]}(u_j)$. Since $\mathcal{M}^{td}$ satisfies **C4** (IoA), we obtain $\mathcal{R}_{[i]}(x) \cap \mathcal{R}^s_{[i]}(u^*) \neq \emptyset$, implying that there exists some $x' \in \mathcal{R}_{[i]}(x) \cap \mathcal{R}^s_{[i]}(u^*)$. It follows from **D2** and statement (2.8) that $x' \notin \mathcal{R}_{\otimes_{Ag}}(w)$, which together with the fact that $u^* \in \mathcal{R}_{\otimes_{Ag}}(w)$ (statement 2.5), implies by criterion u2 of the util function of M^u (Definition 2.15) that $\mathsf{util}(x') < \mathsf{util}(u^*)$. Furthermore, by the construction of M^u, we have $x' \in \mathsf{R}_{[i]}(x) \cap \mathsf{R}^s_{[i]}(u^*)$, $u^* \in \mathsf{R}_{[i]}(u) \cap \mathsf{R}^s_{[i]}(u^*)$ and $\mathsf{util}(x') < \mathsf{util}(u^*)$, which implies the desired claim $\mathsf{R}_{[i]}(u) \not\preceq \mathsf{R}_{[i]}(x)$.

Right-to-left. Assume $\mathsf{M}^u, w \models \otimes_i \varphi$. We reason towards a contradiction by assuming $\mathcal{M}^{td}, w \not\models \otimes_i \varphi$. Hence, there exists a $v \in \mathcal{R}_{\otimes_i}(w)$ such that $\mathcal{M}^{td}, v \not\models \varphi$. Since $\mathcal{M}^{td}$ satisfies **D4** we obtain $\mathcal{R}_{[i]}(v) \subseteq \mathcal{R}_{\otimes_i}(w)$ and hence $\mathcal{R}_{[i]}(v) \not\subseteq \|\varphi\|_{\mathcal{M}^{td}}$. By IH and the construction of M^u, we obtain $\mathsf{R}_{[i]}(v) \not\subseteq \|\varphi\|_{\mathsf{M}^u}$. This fact, in conjunction with the assumption $\mathsf{M}^u, w \models \otimes_i \varphi$, implies that there exists some $\mathsf{R}_{[i]}(z) \subseteq \mathsf{R}_{\Box}(w)$ such that the following hold: (i) $\mathsf{R}_{[i]}(v) \prec \mathsf{R}_{[i]}(z)$, (ii) $\mathsf{R}_{[i]}(z) \subseteq \|\varphi\|_{\mathsf{M}^u}$, and (iii) $\forall \mathsf{R}_{[i]}(x) \subseteq \mathsf{R}_{\Box}(w)$, if $\mathsf{R}_{[i]}(z) \preceq \mathsf{R}_{[i]}(x)$ then $\mathsf{R}_{[i]}(x) \subseteq \|\varphi\|_{\mathsf{M}^u}$.

By Lemma 2.13−(2) and the fact that $\mathsf{R}_{[i]}(z) = \mathcal{R}_{[i]}(z)$, we know that either (a) $\mathcal{R}_{[i]}(z) \subseteq \mathcal{R}_{\otimes_i}(w)$ or (b) $\mathcal{R}_{[i]}(z) \cap \mathcal{R}_{\otimes_i}(w) = \emptyset$ is the case.

Suppose (a) is the case. We know that $\mathsf{R}_{[i]}(v) \prec \mathsf{R}_{[i]}(z)$ and therefore, $\mathsf{R}_{[i]}(z) \not\preceq \mathsf{R}_{[i]}(v)$. Hence, there is a $\mathsf{R}^s_{[i]}(x) \subseteq \mathsf{R}_{\Box}(w)$ with $z^* \in \mathsf{R}_{[i]}(z) \cap \mathsf{R}^s_{[i]}(x)$, and $v^* \in \mathsf{R}_{[i]}(v) \cap \mathsf{R}^s_{[i]}(x)$, such that $\mathsf{util}(v^*) < \mathsf{util}(z^*)$. We also know that $\mathcal{R}_{[i]}(v) \subseteq \mathcal{R}_{\otimes_i}(w)$ and $\mathcal{R}_{[i]}(z) \subseteq \mathcal{R}_{\otimes_i}(w)$ and thus we obtain $z^*, v^* \in \mathcal{R}_{\otimes_i}(w) \cap \mathcal{R}^s_{[i]}(x)$. Consequently, by criterion u3 of the util function of M^u (Definition 2.15) we obtain $\mathsf{util}(v^*) = \mathsf{util}(z^*)$. Contradiction.

Suppose (b) is the case. We know that $\mathsf{R}_{[i]}(v) \prec \mathsf{R}_{[i]}(z)$ and therefore, $\mathsf{R}_{[i]}(z) \not\preceq \mathsf{R}_{[i]}(v)$. Hence, there is a $\mathsf{R}^s_{[i]}(x) \subseteq \mathsf{R}_{\Box}(w)$ with $z^* \in \mathsf{R}_{[i]}(z) \cap \mathsf{R}^s_{[i]}(x), v^* \in \mathsf{R}_{[i]}(v) \cap \mathsf{R}^s_{[i]}(x)$, such that $\mathsf{util}(z^*) \not\leq \mathsf{util}(v^*)$. Then, by criterion u1 of the util function of M^u (Definition 2.15), either (I) $z^* \notin \mathcal{R}^s_{[i]}(x) \setminus \mathcal{R}_{\otimes_i}(w)$ or (II) $v^* \notin \mathcal{R}^s_{[i]}(x) \cap \mathcal{R}_{\otimes_i}(w)$. Suppose (I), since $z^* \in \mathsf{R}^s_{[i]}(x)$ we know that $z^* \in \mathcal{R}^s_{[i]}(x)$ and thus conclude $z^* \in \mathcal{R}_{\otimes_i}(w)$. However, by the initial assumption $\mathcal{R}_{[i]}(z) \cap \mathcal{R}_{\otimes_i}(w) = \emptyset$ we obtain $z^* \notin \mathcal{R}_{\otimes_i}(w)$. Contradiction. Suppose (II), then since $v^* \in \mathcal{R}^s_{[i]}(x)$ we infer $v^* \notin \mathcal{R}_{\otimes_i}(w)$. However, $\mathcal{R}_{[i]}(v) \subseteq \mathcal{R}_{\otimes_i}(w)$. Contradiction. QED

Lemma 2.16. *For each $\varphi \in \mathcal{L}^{td}_n$ we have: $\models_{\mathsf{TUS}_n} \varphi$ implies $\models_{\mathsf{TDS}_n} \varphi$.*

Proof. We prove the claim by contraposition. Assume $\not\models_{\mathsf{TDS}_n} \varphi$. Then, there is a TDS_n-model $\mathcal{M}^{td}$ such that $\mathcal{M}^{td}, w \models \neg\varphi$ for some $w \in \mathcal{W}$. Let M^u be the model obtained from by $\mathcal{M}^{td}$ (Definition 2.15). By Lemma 2.14 we know M^u is a TUS_n model. Last, by Lemma 2.15 we know $\mathsf{M}^u, w \models \neg\varphi$ and, so, $\not\models_{\mathsf{TUS}_n} \varphi$. QED

As an immediate corollary of the above, we know that the Hilbert-style axiomatization of TDS_n is complete with respect to the class of TUS_n-models (Definition 2.13).

Theorem 2.5 (Weak completeness of TUS_n)**.** *For all $\varphi \in \mathcal{L}_n^{td}$, if $\models_{\mathsf{TUS}_n} \varphi$, then $\vdash_{\mathsf{TDS}_n} \varphi$.*

Proof. Follows from Theorem 2.16 together with Theorem 2.3. QED

Moreover, we now know that the two semantic approaches are equivalent.

Theorem 2.6. *For all $\varphi \in \mathcal{L}_n^{td}$, $\models_{\mathsf{TDS}_n} \varphi$, iff $\models_{\mathsf{TUS}_n} \varphi$.*

Proof. Follows directly from Theorems 2.3 and 2.4 together with Lemma 2.16. QED

In fact, since the function util of Definition 2.13 is defined independently of the relations $\mathcal{R}_\mathsf{G}$, $\mathcal{R}_\mathsf{H}$, and $\mathcal{R}_{[Ag]}$, all of the results in this section also hold for DS_n and US_n (including strong completeness).

Corollary 2.3. *For all $\varphi \in \mathcal{L}_n^{d}$, $\models_{\mathsf{DS}_n} \varphi$ iff $\models_{\mathsf{US}_n} \varphi$ iff $\vdash_{\mathsf{DS}_n} \varphi$.*

Remark 2.3. *The atemporal logics US_n and DS_n cannot differentiate between utility functions that are restricted to moments—such as in Definition 2.13—and those functions that assign utilities uniformly to complete histories (i.e., where every world on a history has the same utility). To see this point, consider Murakami's (2005) observation concerning US_n: "[s]ince the formal language [...] contains no operators whose interpretation involves temporal reference [...], and thus from a technical point of view, the temporal relation in utilitarian stit frames can be eliminated when stit formulas and ought formulas are in question" (p.7). In other words, US_n cannot differentiate between multi- and single-moment models, cf. (Balbiani et al., 2008). Since a history of a single-moment model is just a single world, from the perspective of US_n, the two types of utility functions yield the same logic. We further investigate this in the next section.*

2.4 The Limits of Utilities: Temporal Contrary-to-Duty Obligations

So far, we have filled a long-standing gap in the literature by providing a temporal characterization of Deontic STIT logic. Furthermore, in Section 2.3, we showed that utilitarian semantics that assign utilities to moments is equivalent to the relational characterization of the logic TDS_n. Furthermore, we observed that the atemporal characterization of deontic STIT, i.e., DS_n, cannot differentiate between utility functions restricted to moments and those that assign utilities to complete histories (Remark 2.3). The latter extends the results by Murakami (2005), who showed that the atemporal axiomatization of deontic STIT logic cannot distinguish between the following three semantic characterizations: (a) utilitarian STIT for dominance ought (using the reals); (b) utilitarian STIT for optimal ought (using finite-choice models[13]), and (c) the two-valued optimal ought (using binary assignments). Thus, we obtain a class of seven equivalent semantic characterizations with respect to DS_n. This section formally investigates whether these equivalences are preserved for the logic TDS_n. We are now in the position to answer an open question posed by Murakami (2005) and investigate "how various operators for deontic notions behave and interact in a temporal structure" (p.5). In particular, we investigate the following claim by Horty (2001):

> [B]ecause the utilitarian setting allows us to handle reparational oughts [i.e., CTD obligations[14]] while maintaining a uniform assignment of values to histories, and because such an assignment seems more natural—we build this uniformity constraint into our definition of the utilitarian framework. (p.41)

To understand the reasons given by Horty in the above quote, consider the utility assignment in Figure 2.2 on page 36. At moment $\mathcal{R}_\square(\omega^\alpha)$, John and Paul are both obliged to try to work things out. If the two agents act according to their duty, this eventually yields a utility of at least 3 at moment $\mathcal{R}_\square(v_i)$. In contrast, if both violate their obligation and choose not to work it out—i.e., the choices $\{\omega^u, \omega^x\}$, respectively $\{\omega^z, \omega^x\}$—at $\mathcal{R}_\square(\omega^\alpha)$, they arrive at a sub-ideal moment $\mathcal{R}_\square(x_i)$ where the CTD obligation consists in "getting a little help from their friends". At this violation state $\mathcal{R}_\square(x_i)$, the CTD obligation assures the maximum utility of 2 if observed. Although this utility is at $\mathcal{R}_\square(x_i)$ the highest, it is strictly less than the assured utility of 3 at $\mathcal{R}_\square(v_i)$ resulting from John and Paul fulfilling their initial obligation to try to work it out together. Hence, the use of utilities assigned to histories naturally represents that the CTD situation in $\mathcal{R}_\square(x_i)$ is strictly less ideal than the according to duty situation in $\mathcal{R}_\square(v_i)$. We refer to Horty (2001, Ch.3) for a more detailed discussion.

In this section, we assess the claim made by Horty on page 64 above. Our conclusion will be twofold: first, the two-valued utility function—e.g., investigated in (Murakami, 2005)—causes problems concerning CTD scenarios. Second, the observed equivalence between different utility functions in the context of DS_n is not preserved, i.e., TDS_n is incomplete with respect to two-valued utility functions assigning utilities to complete histories. Consequently, our analysis provides a formal argument why only the real/natural numbers are suitable for Temporal Utilitarian STIT logic. With this, we address Objective 4.

2.4.1 Utilities Assigned to Histories

We first define some preliminaries. In what follows, we use util^m and util^h to differentiate between utility functions that are restricted to moments, respectively histories (the utility function in Definition 2.13 is a util^m function. For each world w, we define a *history* as

[13] Following Horty (2001), the optimal ought characterizes dominance act utilitarianism proper. The definition of (a) is a generalization of (b), the latter which is characterized by finite-choice frames.

[14] Like Governatori and Rotolo (2006), Horty (2001) calls obligations that arise through the violation of previous obligations 'reparational oughts' (cf. CTD-reasoning in Chapter 1).

the set $\mathsf{h}(w) = \mathcal{R}_{\mathsf{H}}(w) \cup \{w\} \cup \mathcal{R}_{\mathsf{G}}(w)$. It can be straightforwardly observed that each world is a member of exactly one history and that $\mathcal{R}_{\mathsf{G}}$ forms a strict linear order on $\mathsf{h}(w)$ (Definition 2.4). We refer to temporal utilitarian STIT frames that employ util^h functions as hTUS_n-models.

Definition 2.16 (Frames and Models for hTUS_n)**.** *A relational* Temporal Utilitarian STIT frame *restricted to* histories *(for short,* hTUS_n*-frame) is a tuple* $\mathfrak{F} = \langle W, \mathcal{R}_\Box, \{\mathcal{R}_{[i]} | i \in Ag\}, \mathsf{util}^h, \mathcal{R}_{[Ag]}, \mathcal{R}_{\mathsf{G}}, \mathcal{R}_{\mathsf{H}}\rangle$*, where* $\mathfrak{F}$ *satisfies conditions* ***C1****–****C6*** *and* ***T1****–****T6*** *of Definition 2.4 together with the following:*

U1 $\mathsf{util}^h : W \mapsto \mathbb{N}$ *is a utility function assigning each world a natural number;*

U2 *For all* $w, v \in W$ *such that* $v \in \mathsf{h}(w)$*, we have* $\mathsf{util}^h(w) = \mathsf{util}^h(v)$*.*

A hTUS_n*-model is a tuple* $\mathfrak{M} = \langle \mathfrak{F}, V\rangle$ *where* $\mathfrak{F}$ *is a* hTUS_n*-frame and* V *is a valuation function assigning propositional atoms to subsets of* W*, i.e.,* $V\colon \mathsf{Atoms} \mapsto \wp(W)$*.*

The main difference between hTUS_n- and TUS_n-models is that the utility function is restricted to assigning the same utility to all worlds belonging to the same history, i.e., **U2**. In other words, we may say that the utility is assigned to the history itself. The satisfaction of a $\mathcal{L}_n^{td}$ formulae on a hTUS_n-model is defined as in Definition 2.14.

2.4.2 Example: A Temporal Contrary-to-Duty Scenario

In what follows, we model a CTD scenario using *deliberative* obligations. There is a strong conceptual connection between deliberative obligations and CTD reasoning: both require that obligations can be violated (Governatori and Rotolo, 2006). A deliberative obligation is defined as:

$$\otimes_i^d \varphi := \otimes_i \varphi \land \Diamond \neg \varphi$$

From the point of view of agency, deliberative obligations ensure that the agent's choices are somehow influential to whether the obligation is satisfied or violated. Tautological obligations are vacuously satisfied and, for that reason, do not classify as deliberative. It can be straightforwardly observed that $\otimes_i^d$ forces *at least two* available choices for the agent involved. CTD reasoning likewise assumes the violability of obligations: it deals with those obligations that hold in scenarios where another obligation is violated. We discuss an explicitly temporal CTD scenario (see Chapter 1 for a general introduction).[15]

Example 2.4 (A Temporal CTD Scenario)**.** *The scenario presented in Example 2.2 is a temporal CTD scenario. John and Paul are obliged to try to work it out. If*

[15] The standard approach is to take CTD reasoning as an atemporal problem (Hilpinen and McNamara, 2013). Although some CTD scenarios can be adequately addressed in an explicitly temporal language, it does not provide a uniform solution to CTD reasoning, e.g., see Prakken and Sergot (1996). Since our aim in this chapter is to develop an explicitly temporal deontic STIT logic, we focus on temporal CTD scenarios. See Chapter 6 for a discussion of atemporal CTD reasoning.

one of them fails to comply, a CTD obligation ensues, requiring both agents to get some help from their friends. Furthermore, if both comply with their initial duty, they become obliged to thank each another. We assume that all involved obligations are deliberative obligations. *The scenario is graphically presented in Figure 2.2 on page 37 (see Example 2.2 for an explanation of the figure). Observe that of the four successor moments $\mathcal{R}_{\Box}(v_i), \mathcal{R}_{\Box}(u_i), \mathcal{R}_{\Box}(x_i)$, and $\mathcal{R}_{\Box}(z_i)$ of $\mathcal{R}_{\Box}(\omega^{\alpha})$, three depict CTD moments. The scenario consists of the following three formulae:*

E1. $\otimes_j^d \mathtt{try\_j} \wedge \otimes_p^d \mathtt{try\_p}$;

E2. $\Box(\mathtt{work\_it\_out} \rightarrow ([j]\mathtt{try\_j} \wedge [p]\mathtt{try\_p})) \wedge \Diamond \mathtt{work\_it\_out}$;

E3. $\Box(\mathtt{work\_it\_out} \rightarrow \mathsf{F}(\otimes_j^d \mathtt{thank\_j} \wedge \otimes_p^d \mathtt{thank\_p}))$;

E4. $\Box(\neg \mathtt{work\_it\_out} \rightarrow \mathsf{F}(\otimes_j^d \mathtt{help\_j} \wedge \otimes_p^d \mathtt{help\_p}))$.

We point out that the irreflexivity of TDS_n- and TUS_n-frames ensures that the CTD obligation expressed in E3 *becomes effective strictly after the violation has taken place. Henceforth, we use $\Sigma_{\mathsf{Ex}} = \{$E1, E2, E3,E4$\}$ to refer to the above scenario.*

Figure 2.2 provides a consistent representation of Σ_{Ex} in the logics TDS_n, TUS_n, and hTUS_n, where $\mathfrak{M}, \omega^{\alpha} \models \Sigma_{\mathsf{Ex}}$ for each $\alpha \in \{v, u, z, x\}$. Recall that by seriality and irreflexivity, the underlying branching time frame is infinite; therefore, the model graphically represented in this figure is only partial. Below, we define the model formally for each semantics.

The Models Satisfying Σ_{Ex}. Let $\mathsf{Agents} = \{j, p\}$. We recall that TDS_n-, TUS_n-, and hTUS_n-models only differ on how $\otimes_i$ is interpreted. Hence, we can uniformly define the non-deontic elements of these models. We do this first. We start by defining the infinite set of worlds (forming our domain) together with all moments and choices:

- $W = W_{\omega} \cup W_v \cup W_u \cup W_z \cup W_x \cup W'$, where $W_{\omega} = \{w_i^{\alpha} \mid \alpha \in \{v, u, z, x\}$ and $i \in \{1, 2, 3, 4\}\}$ and $W_{\alpha} = \{\alpha_i \mid i \in \{1, 2, 3, 4\}\}$ for each $\alpha \in \{v, u, z, x\}$. Let $W' = \{\alpha_i^j \mid \alpha \in \{v, u, z, x, \}, i \in \{1, 2, 3, 4\}$ and $j \in \mathbb{N}\}$.
- $\mathcal{R}_{\Box}(w_i^{\alpha}) = W_{\omega}$ for each $w_i^{\alpha} \in W_{\omega}$, $\mathcal{R}_{\Box}(\alpha_i) = W_{\alpha}$ for $\alpha \in \{v, u, z, x\}$ and $i \in \{1, 2, 3, 4\}$, and $\mathcal{R}_{\Box}(\alpha) = \{\alpha\}$ for each $\alpha \in W'$.
- $\mathcal{R}_{[j]}(\omega^{\alpha}) = \{w_i^v \mid 1 \leq i \leq 4\} \cup \{w_i^z \mid 1 \leq i \leq 4\}$ with $\alpha \in \{v, z\}$ and $\mathcal{R}_{[j]}(\omega^{\alpha}) = \{w_i^u \mid 1 \leq i \leq 4\} \cup \{w_i^x \mid 1 \leq i \leq 4\}$ with $\alpha \in \{u, x\}$. $\mathcal{R}_{[j]}(\alpha_i) = \{\alpha_1, \alpha_3\}$ and $\mathcal{R}_{[j]}(\alpha_k) = \{\alpha_2, \alpha_4\}$ for $\alpha \in \{v, u, z, x\}$ and $i \in \{1, 3\}$ and $k \in \{2, 4\}$.
- $\mathcal{R}_{[p]}(\omega^{\alpha}) = \{w_i^v \mid 1 \leq i \leq 4\} \cup \{w_i^u \mid 1 \leq i \leq 4\}$ with $\alpha \in \{v, u\}$ and $\mathcal{R}_{[p]}(\omega^{\alpha}) = \{w_i^z \mid 1 \leq i \leq 4\} \cup \{w_i^x \mid 1 \leq i \leq 4\}$ with $\alpha \in \{z, x\}$. $\mathcal{R}_{[p]}(\alpha_i) = \{\alpha_1, \alpha_2\}$ and $\mathcal{R}_{[p]}(\alpha_k) = \{\alpha_3, \alpha_4\}$ for $\alpha \in \{v, u, z, x\}$ and $i \in \{1, 2\}$ and $k \in \{3, 4\}$.

- $\mathcal{R}_{[j]}(\alpha) = \mathcal{R}_{[p]}(\alpha) = \{\alpha\}$ for each $\alpha \in W'$.[16]
- For all $w \in W$ we define $\mathcal{R}_{[Ag]}(w) = \bigcap_{i \in \mathsf{Agents}} \mathcal{R}_{[i]}(w)$.

We define the temporal structure over W as follows:

- First, let $\mathcal{R}^*_{\mathsf{G}} = \{(w_i^\alpha, \alpha_i) \mid \alpha \in \{v,u,z,x\}$ and $i \in \{1,2,3,4\}\} \cup \{(\alpha_i, \alpha_i^1) \mid \alpha \in \{v,u,z,x\}, i \in \{1,2,3,4\}\} \cup \{(\alpha_i^j, \alpha_i^{j+1}) \mid \alpha \in \{v,u,z,x\}, i \in \{1,2,3,4\}$, and $j \in \mathbb{N}\}$. We define $\mathcal{R}_{\mathsf{G}}$ as the transitive closure of $\mathcal{R}^*_{\mathsf{G}}$.
- $\mathcal{R}_{\mathsf{H}} = \{(\alpha, \beta) \mid (\beta, \alpha) \in \mathcal{R}_{\mathsf{G}}\}$.

Let $\mathsf{Atoms} = \{\mathtt{work\_it\_out}, \mathtt{try\_j}, \mathtt{try\_p}, \mathtt{help\_j}, \mathtt{help\_p}, \mathtt{thank\_j}, \mathtt{thank\_p}\}$, the valuation of atoms is defined as:

- $V(\mathtt{work\_it\_out}) = \{w_i^v \mid 1 \leq i \leq 4\}$, $V(\mathtt{try\_j}) = \{w_i^\alpha \mid \alpha \in \{v,z\}$ and $1 \leq i \leq 4\}$, $V(\mathtt{try\_p}) = \{w_i^\alpha \mid \alpha \in \{v,u\}$ and $1 \leq i \leq 4\}$, $V(\mathtt{help\_j}) = \{\alpha_i \mid \alpha \in \{u,z,x\}$ and $i \in \{1,3\}\ \}$, $V(\mathtt{help\_p}) = \{\alpha_i \mid \alpha \in \{u,z,x\}$ and $i \in \{1,2\}\ \}$, $V(\mathtt{thank\_j}) = \{v_i \mid i \in \{1,3\}\ \}$, $V(\mathtt{thank\_p}) = \{v_i \mid i \in \{1,2\}\ \}$.

Furthermore, each util^h is a util^m function by Definition 2.16. Therefore, it suffices to define a TDS_n- and hTUS_n-model (the latter being a TUS_n model too).

- For the TDS_n interpretation of Σ_{Ex}:

  $\mathcal{R}_{\otimes_j} = W' \cup \{w_i^\alpha \mid \alpha \in \{v,z\}$ and $1 \leq i \leq 4\} \cup \{\alpha_i \mid \alpha \in \{v,u,z,x\}$ and $i \in \{1,3\}\}$.
  $\mathcal{R}_{\otimes_p} = W' \cup \{w_i^\alpha \mid \alpha \in \{v,u\}$ and $1 \leq i \leq 4\} \cup \{\alpha_i \mid \alpha \in \{v,u,z,x\}$ and $i \in \{1,2\}\}$.

- For the hTUS_n interpretation of Σ_{Ex}:

  For each $\alpha \in \{\alpha \in \mathsf{h}(v_i) \mid 1 \leq i \leq 3\}$, $\mathsf{util}^h(\alpha) = 4$. For each $\alpha \in \mathsf{h}(v_4)$, $\mathsf{util}^h(\alpha) = 3$. For each $\alpha \in \{\alpha \in \mathsf{h}(\beta_i) \mid \beta \in \{u,z,x\}$ and $1 \leq i \leq 3\}$, $\mathsf{util}^h(\alpha) = 2$. For each $\alpha \in \{\alpha \in \mathsf{h}(\beta_4) \mid \beta \in \{u,z,x\}\}$, $\mathsf{util}^h(\alpha) = 1$.

It can be straightforwardly checked that the above defines a TDS_n-, a TUS_n-, and a hTUS_n-model. We point out that util^h satisfies both **U1** and **U2**, and thus it is also a util^m function. Furthermore, both utility functions range over the natural numbers.

One can likewise define a *two-valued* utility function util^m—ranging over $\{1,0\}$—for the TUS_n-model, while preserving satisfiability of Σ_{Ex}. Namely, for each individual moment, one assigns 1 to each world in an optimal choice and 0 otherwise. Intuitively, the reason is that we can reset the utility assignment for each moment occurring in the tree. We now discuss the behavior of binary functions in the context of util^h.

[16] This clause defines the infinite continuation of each history with single-world moments.

2.4.3 The Problem with Two-Valued Utility Functions

Murakami (2005) showed that the axiomatization of atemporal deontic STIT logic (cf. DS_n) is sound and complete for various types of utility functions (see page 64). Hence, DS_n is not expressive enough to distinguish between these functions. This includes the two-valued utility function, assigning either 0 or 1 to each *history*. In this last section, we show that the *temporal* TDS_n is *incomplete* with respect to the class of hTUS_n-frames using a binary util^h function.

We start with a general observation. There are two limiting cases in the two-valued approach to hTUS_n: (†) at a moment where all histories passing through the moment have a utility of 1 every obligation becomes vacuously satisfied by definition—in such a scenario, we have $\otimes_i\varphi$ iff $\Box\varphi$—and every choice for each agent will ensure all optimal outcomes. The same reasoning applies to moments where (‡) all histories passing through that moment have a utility of 0.[17] By assigning utilities to complete histories (Horty, 2001), in the case of (†) and (‡), each obligation holding at a future moment will be vacuously satisfied. Namely, as one progresses in time, the set of histories passing through a moment can only decrease or stay the same, and therefore the assigned utilities will remain 1 for each future moment in the case of (†) and will remain 0 in the case of (‡). That all obligations are vacuously satisfied at such moments means that no obligation can be violated. This also implies that at such moments there are no deliberative obligations possible. Consequently, contrary-to-duty reasoning becomes impossible at these moments because CTD scenarios require the possibility of violating an obligation.

As argued above, in order to reason with CTD scenarios in *temporal* utilitarian STIT logics, we need to ensure that obligations can be violated. For that reason, we consider deliberative obligations. For an obligation $\otimes_i^d\varphi$ to hold, there exists a choice that does not guarantee φ—i.e., $\langle i\rangle\neg\varphi$—and, by Definition 2.14, the latter choice must be *strictly dominated* by φ choices. In the binary setting, this means that for all deontically optimal choices, there is at least one outcome with a strictly higher utility: in the case of a two-valued util^h function, this utility must be 1. Unfortunately, this fact has the drawback that at such moments *at least* one of the following two statements holds:

S1 All histories in the intersection of all agents acting in accordance with their duty *have a utility of* 1;

S2 All histories in the intersection of all agents violating their duty *have a utility of* 0.

Relative to statements S1 and S2, reconsider the scenario of John and Paul in a *two-agents two-choices* setting where $\otimes_j^d\mathtt{try\_j} \wedge \otimes_p^d\mathtt{try\_p}$. Figure 2.3 illustrates the only three two-valued utility assignments possible for satisfying these two obligations. We argue that all three assignments are problematic. In what follows, with the “impossibility of future CTD scenarios”, we mean that all (future) obligations will be vacuously satisfied from

[17] The observation also applies to moments where all intersections of choices of agents contain *both* 1 and 0. Nevertheless, this observation is not needed for the argument made in this section.

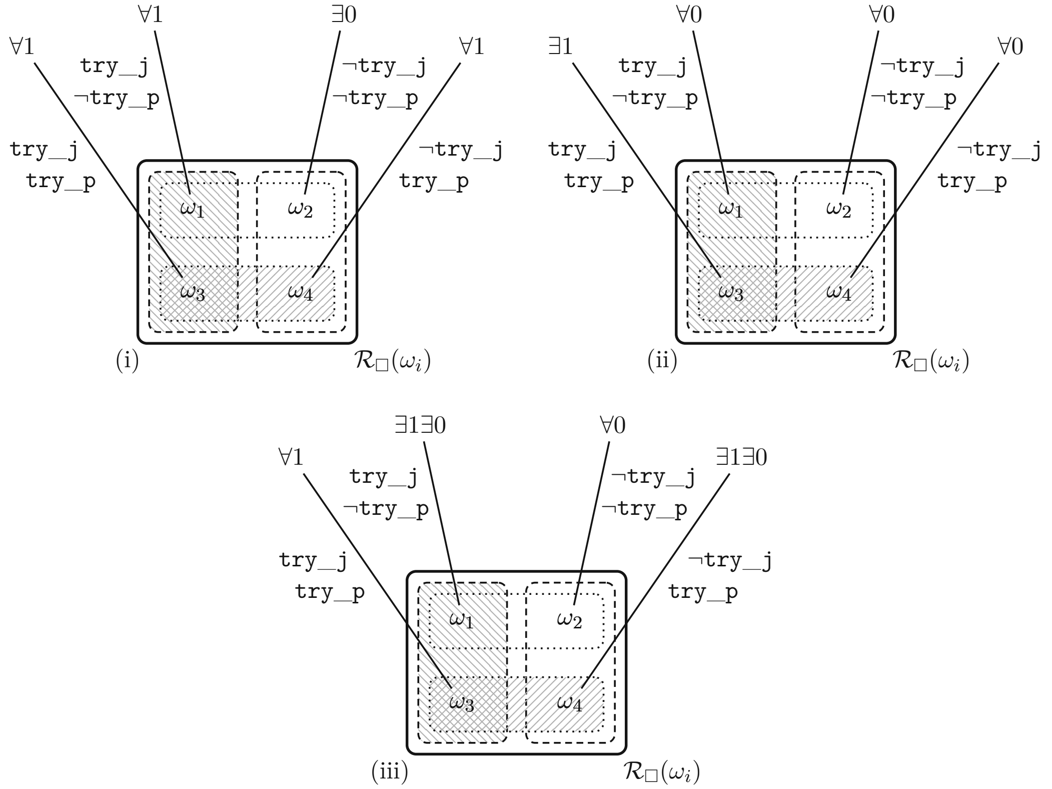


Figure 2.3: The only three scenarios (i)-(iii) for which $[j]^d$`try_j` and $[p]^d$`try_p` are satisfiable on a two-agent two-choice TUS_n-model. ω_i denotes an arbitrary non-empty set of worlds, for $1 \leq i \leq 4$. Choices of j are vertically presented, and those of p horizontally. The symbol $\forall k$ at ω_i with $i \in \{1,2,3,4\}$ means that every history $\mathsf{h}(\omega_i)$ is assigned value k, and $\exists k$ at ω_i means that some history $\mathsf{h}(\omega_i)$ going through the choice intersection is assigned k, for $k \in \{0,1\}$. Deontically optimal choices are shaded and darker shaded when overlapping. At all $\forall k$ outcomes with $k \in \{0,1\}$, all obligations will be vacuously satisfied forever onward, and so CTD reasoning becomes impossible.

that moment onward. Sub-figure (i) implies the impossibility of future CTD reasoning in all cases in which at least one agent satisfies the obligation, i.e., in those cases, all future obligations will be vacuously satisfied. Sub-figure (ii) implies that there are no future CTD scenarios possible in each case witnessing at least one agent violating his obligation. Last, sub-figure (iii) indicates that future CTD scenarios can only occur in those cases when one of the agents satisfies his obligation *provided* that the other is in violation. One may check the exhaustiveness of these scenarios by inspecting the semantic interpretation of $\otimes^d_j$`try_j` $\wedge$ $\otimes^d_p$`try_p`. For all other utility assignments, contradictions ensue. The above analysis generalizes to an arbitrary number of agents with an arbitrary number of choices. (Observe that Figure 2.3 satisfies S1 and S2.)

In Figure 2.3, we see that for each of the three scenarios in at least one future moment, deliberative obligations are impossible, i.e., whenever the future moment satisfies $\forall 1$ or $\forall 0$. Consequently, the temporal CTD scenario of Example 2.4 cannot be satisfied on any hTUS_n-model adopting a two-valued utility function. In other words, Σ_{Ex} is unsatisfiable. However, Σ_{Ex} is satisfiable with respect to the class of TDS_n- and TUS_n-models, as shown on page 66.

Theorem 2.7. *The Hilbert-style axiomatization of* TDS_n *is incomplete with respect to the class of* hTUS_n*-frames with a* util^h *function mapping histories to* $\{1, 0\}$

Proof. By Theorem 2.3, TDS_n is sound and complete with respect to the class of TDS_n-frames. Recall $\Sigma_{\mathsf{Ex}} = \{\mathsf{E1},\mathsf{E2},\mathsf{E3},\mathsf{E4}\}$. The formula $\bigwedge \Sigma_{\mathsf{Ex}}$ is TDS_n-satisfiable (page 66). For any hTUS_n-model with a binary utility function, we know that if $\mathsf{E1}$ is satisfiable either S1 or S2 holds. Therefore, at least one future moment contains only vacuously satisfied choices. Thus, either $\mathsf{E3}$ or $\mathsf{E4}$ cannot be satisfied, and so $\bigwedge \Sigma_{\mathsf{Ex}}$ is unsatisfiable. Hence, $\neg \bigwedge \Sigma_{\mathsf{Ex}}$ is a valid formula of two-valued hTUS_n-models and, since $\neg \bigwedge \Sigma_{\mathsf{Ex}}$ is not TDS_n-derivable, TDS_n is incomplete with respect to the class of two-valued hTUS_n-frames. QED

How should we interpret the incompleteness result of Theorem 2.7? Murakami (2005) showed that atemporal deontic STIT logics are indifferent with respect to utility assignments from $\{0, 1\}$ and $\mathbb{N}$ (or $\mathbb{R}$ for that matter). Although utility functions relative to moments come with their own challenges (Horty, 2001), this section demonstrated that two-valued utility functions are unsuitable for deliberative agency in the context of explicit temporal reasoning. The results in this section provide strong support for adopting natural or real numbers for temporal utilitarian STIT logics, e.g., as proposed by Horty (2001).

2.5 Related Work and Future Research

Decidability. Decidability of STIT logics has been extensively investigated. Basic STIT logic was shown decidable by Xu (1994b). Xu (1994a) also showed the decidability of a deliberative STIT logic that takes the deliberative STIT modality as a primitive instead of a defined modality. An alternative approach to the decidability of these logics was given by Balbiani et al. (2008), who also showed that the settledness modality $\Box$ can be omitted from the language by defining it in terms of choice operators. A proof-theoretic decidability result—including proof-search algorithms and automated counter-model extraction—was developed for basic STIT logic in (Lyon and Berkel, 2019). A similar system was introduced by Negri and Pavlović (2021). Group STIT logics, where choice is considered not only in relation to individual agents but also in relation to arbitrary groups of agents, was shown undecidable by Herzig and Schwarzentruber (2008). Furthermore, Murakami (2005) provided a semantic proof showing that Utilitarian STIT

logic is decidable. A proof-theoretic proof of the decidability of the equivalent logic DS_n—including proof-search algorithms and automated counter-model extraction—was given by Lyon (2021).[18] Last, Ciuni and Lorini (2018) investigate decidability of various temporal extensions of basic STIT logic using different temporal semantics. The decidability of the Temporal STIT logic (Lorini, 2013) on which TDS_n is based is still an open question. In light of the above, the following proves an interesting future research direction:

Open question 2.1. *Is* TDS_n *decidable? If not, is there an alternative axiomatization of Temporal Deontic* STIT *logic, in the spirit of (Ciuni and Lorini, 2018), which is decidable?*

Other Temporal STIT Logics. The logic TDS_n is based on the temporal axiomatization of BT+AC frames as developed by Lorini (2013). However, this is not the only temporal STIT logic in the literature. A central feature of the basic STIT operator $[i]$ is that it is instantaneous, i.e., referring to choice at the present moment. The logic of XSTIT contains a non-instantaneous STIT-operator explicitly affecting next states. Introduced by Broersen (2011b), the logic is motivated by the observation that affecting next states is a central aspect of agency in computer science. Moreover, extensions of the logic XSTIT have been employed to investigate the concepts of purposeful and voluntary acts and their relation to different levels of legal culpability (Broersen, 2011a). The logic was initially proposed for a two-dimensional semantics referring to both states and histories. An alternative semantic characterization of XSTIT—using relational semantics—was provided in (van Berkel and Lyon, 2019b); there, it was shown that the two semantic approaches are equivalent. Sequent-style proof systems for temporal STIT logic and XSTIT logic were likewise given in (van Berkel and Lyon, 2019b). Next, the initial STIT operator proposed in the seminal work of Belnap and Perloff (1988) is also inherently temporal. The operator is called the Achievement STIT (for short, ASTIT). The main characteristic of this logic is that it refers to both the past and alternative courses of events. In brief, an ASTIT formula $[i]^a\varphi$ expresses that “through a choice in the past A holds at the present moment (and was guaranteed to hold), even though the agent’s alternative choices at that moment would not have ensured A (after passing of the same interval of time)” (where A is a formula). The logic was shown complete by Xu (1995).

Epistemic extensions of temporal STIT logics were introduced and analyzed by Broersen (2011a) and Broersen (2011b). Furthermore, Broersen (2011a) discusses deontic modalities in the context of epistemic XSTIT. There, the obligation operator is taken as a defined operator in the Andersonian tradition (Anderson and Moore, 1957), i.e., obligations are reduced to choices leading to sanctions (or violations). Lorini (2013) discusses normative concepts in the context of Temporal STIT logic via adopting an Andersonian reduction of obligations to choices and violations. In particular, the reduction was adopted to reason

[18] Alternative proof systems for STIT logic are provided by Arkoudas et al. (2005), who introduce a natural deduction system for a deontic STIT logic (without soundness and completeness results), and by Olkhovikov and Wansing (2018) and Wansing (2006), who propose tableaux systems for multi-agent deliberative STIT logics.

about commitments. A similar reductionist approach was adopted by Bartha (1993) and Xu (2015) in atemporal STIT settings. We refer to Chapter 4 and Chapter 5 for an extensive discussion of Andersonian approaches to deontic logics.

It remains to be determined whether TDS_n is sound and complete with respect to Temporal Utilitarian STIT semantics employing utility functions assigning naturals to complete histories. Although two-valued utility functions yielded the incompleteness result in Theorem 2.7, we conjecture that the following holds true:

Conjecture 2.1. *The logic* TDS_n *is weakly complete for the class of* TUS_n*-frames.*

Last, this chapter did not deal with important themes such as validity time of obligation versus reference time of obligation, and maintenance versus achievement obligations (Broersen and Torre, 2011). The developed Temporal Deontic STIT logic may also be used in future work to investigate these topics.

Conditional Obligations for STIT logics. The deontic modalities used in this chapter are monadic, i.e., unary modal operators. Several extensions of atemporal STIT have been introduced for dealing with *conditional oughts*. We discuss these briefly here. The earliest account of conditional obligations is given by Bartha (1993). There, different formalizations of conditional obligations are discussed in light of deontic paradoxes, such as CTD scenarios. The conditional is interpreted with the use of material implication. Horty (2001) introduces a notion of choice conditioned on a proposition. The dominance relation over choices (cf. Section 2.3) can also be conditioned on such a proposition. Subsequently, a ternary modality $\mathcal{O}([i]\varphi/\psi)$ is introduced, roughly expressing that "under the condition that ψ is the case, the agent i ought to see to it that φ holds". Last, Sun and Baniasadi (2014) extend group STIT logic (where choice ranges over arbitrary groups of agents) with a monadic and a dyadic group obligation. Their account is based on Utilitarian STIT semantics. The resulting logic is applied to the Miners Paradox (Kolodny and MacFarlane, 2010). Also, see the work of Abarca and Broersen (2019) for an analysis of the Miners Paradox in the context of epistemic deontic STIT logic.

Since the adaptation of conditional obligations has proven to be a fruitful approach for dealing with deontic paradoxes in general and (atemporal) CTD scenarios in particular (Hilpinen and McNamara, 2013), it is left to future work to analyze the logical behavior of such obligations in the context of an explicitly temporal STIT setting.

Open question 2.2. *What are the logical properties of expanding* TDS_n *with conditional obligations?*

The Logic of Bringing it About That. A formalism similar to STIT is the logic of 'bringing it about that' by Elgesem (1997). As for STIT, Elgesem's logic represents agency through a canonical form, namely, "bringing about that" (which is considered by Belnap and Perloff (1988) as a synonym for seeing to it that). Elgesem stresses problematic aspects of using normal modal operators for capturing agency—such as those employed in basic STIT logic—due to, e.g., necessitation. Goal-directed behavior is a central notion for Elgesem's theory of agency. This behavior ensues when the agent causally contributes to attaining a desired result, the latter of which must be non-trivial and non-accidental. Furthermore, it is a counterfactual notion ensuring that the result is due to the agent's capacity and not an accidental by-product. Non-triviality and non-accidentality warrant the use of non-normal modal operators (Chellas, 1980). Elgesem's (1997) logic contains various primitive non-normal modal operators: e.g., the agent-dependent "Does"-modality. The logic (extended with coalitions) was shown to be sound and complete with respect to bi-neighborhood semantics and hyper-sequent systems (Dalmonte et al., 2021). We point out that in Chapter 3, we provide a non-normal modal characterization of the deontic modality $\otimes_i$ in the context of STIT for reasons similar to those in (Chellas, 1980; Elgesem, 1997).

* * *

In this chapter, we filled a long-standing gap in the STIT literature by providing a sound and weakly complete Temporal Deontic STIT logic (Objective 1). We showed that this STIT logic can be semantically characterized using only relational semantics, i.e., bypassing both the traditional BT+AC semantics and the utilitarian STIT semantics (Objective 2). We showed how the resulting semantics can be truth-preservingly transformed into the traditional utilitarian STIT semantics of dominance ought (Horty, 2001) (Objective 3). What is more, we formally investigated the logical consequences of adopting an explicitly temporal language in the context of deontic STIT and proved that the two-valued function, ranging over histories, is unsuitable for temporal CTD reasoning, yielding incompleteness with respect to TDS_n (Objective 4).

CHAPTER 3

Ought Implies Can

The fields of moral philosophy and deontic logic gave rise to various *metaethical principles*. Metaethical principles are requirements that any appropriate ethical theory must *ideally* satisfy. They are principles of a higher generality than the principles and rules within a given ethical theory. Intuitively, one may differentiate the two as follows: On the one hand, particular ethical rules, such as "one should not lie", are about the normative status of specific behavior. They are action-guiding and restricted to particular action-types (McConnell, 1985). Metaethical principles, on the other hand, such as "an ethical theory must be consistent", apply independently of a given action-type and are supposed to hold for any ethical theory.[1] Their generality puts them on the same level as axiom schemes. Following McConnell (1985), metaethical principles serve as preconditions that ethical theories should ideally satisfy: "[i]t is when a view [ethical theory] fails to satisfy several (or many) such conditions that we begin to feel confident placing it outside the realm of the moral" (p.307). Examples of metaethical principles that play a central role in philosophy and the logical analysis of normative reasoning are: "no vacuous obligations" (von Wright, 1951), "deontic contingency" (Anderson and Moore, 1957), "deontic consistency" (Marcus, 1980), "(im)possibility of deontic dilemmas" (Conee, 1982), and the principle of "alternate possibilities" (Copp, 2017). Yet, the most prevalent and extensively discussed metaethical principle is *"Ought implies Can"*.

This chapter is about Ought implies Can (OiC, for short). In its general formulation, the principle stipulates that what ought to be done, *can be done*. One of the allures of OiC is that it releases agents from alleged duties that are impossible, strenuous, or over-demanding (Dahl, 1974; McConnell, 1989). To see this point, consider OiC in contraposition: "what cannot be done, an agent is not obliged to do." In other words, OiC delimits the possible actions to which an agent can be normatively bound. It ensures this by taking into account the agent and the circumstances in which the agent finds

[1] For a discussion of the normative status of metaethical principles see (McConnell, 1985).

herself when reasoning about obligations and norms. Hence, inferences about obligations are influenced by what ‘can be done’.

Unfortunately, there is no clear consensus on the philosophical and logical interpretation of OiC. The principle has a long history within moral philosophy and can be traced back to, for example, Aristotle (*The Nicomachean Ethics*, translated by Ameriks and Clarke 2000, VII-3), and to the ancient Roman legal principle “*impossibilium nulla obligatio est*” (Vranas, 2007). Usually, OiC is accredited to the renowned philosopher Immanuel Kant. For instance, in the Critique of Pure Reason, Kant writes that “of course the action must be possible under natural conditions if the ought is directed to it” (translated by Guyer and Wood 1998, A548/B576). While earlier thinkers such as Aristotle and Kant only discussed OiC implicitly, it became an explicit subject of investigation in the twentieth century. Aside from debates on whether OiC should be adopted at all (Graham, 2011; Saka, 2000), most works are about which *reading* of the principle should be endorsed. In particular, most discussions revolve around the right interpretation of ‘can’. Determining the right interpretation of ‘can’ is crucial for systems that adopt OiC since it influences the degree to which an agent can be burdened with and relieved from duties. Notable positions have been taken up by Hintikka (1970), Lemmon (1962), Stocker (1971), von Wright (1963a), and, more recently, Vranas (2007).

The central aim of this chapter is to enhance our understanding of OiC using tools from formal logic. We focus on frequently recurring readings from authors that are—in our opinion—central to the debate. Despite the apparent relationships between some of the considered OiC readings, a precise taxonomy of their logical interdependencies is only achievable through a formal investigation of their corresponding logics. Such a logical taxonomy is still missing. Although OiC is one of those properties commonly taken as ‘undisputed’ in the field (van der Torre, 1997) there is a severe discrepancy between the formal treatment of OiC and its philosophical counterpart, which it aims to model. This chapter extends the preliminary results obtained in (van Berkel and Lyon, 2021) and fills this gap. To better understand OiC, we develop deontic logics for various OiC interpretations. In particular, we employ the formalism of STIT (Belnap and Perloff, 1988) since it allows us to model obligations and various agential concepts such as ability.

Objective 1. *Develop sound and complete Deontic* STIT *logics for each prominent reading of OiC from the philosophical literature.*

The upshot of employing logical methods is that we can formally determine which interpretations of OiC logically imply others and which readings are logically independent.

Objective 2. *Use the developed Deontic* STIT *logics to determine the logical relations between the various readings of OiC, thus yielding a formal taxonomy of OiC.*

Additionally, we are interested in using the obtained OiC logics to acquire a better understanding of the relations between other metaethical principles and OiC.

Objective 3. *Employ the developed logics for OiC and determine the logical relations to other metaethical principles in the literature.*

Last, we are interested in reasoning with the obtained OiC logics and determining how different interpretations of 'can' influence inferences about obligations.

Objective 4. *Enhance the developed OiC logics with reasoning principles that take into account what 'can be done' in drawing inferences about obligations.*

Contributions. In this chapter, we address these four objectives by introducing a class of logics for the analysis of Ought-implies-Can in STIT. We refer to these logics as OS_n (with n referring to the number of agents in the language). We briefly discuss the five main contributions made in this chapter.

First, we discuss, compare, and formalize ten OiC principles collected via an extensive analysis of the philosophical literature (Section 3.1). To the best of our knowledge, such a classification of OiC principles is novel.

Second, the intrinsically agentive setting provided by the STIT paradigm enables us to conduct a fine-grained analysis of the various renditions of OiC. The traditional, utilitarian approach to deontic STIT logic by Horty (2001) enforces specific properties on its obligation operators, which includes an axiom for Ought implies Can (see axiom A11 of DS_n in Chapter 2). However, most philosophical readings of OiC are either weaker or stronger than the OiC principle of traditional deontic STIT logic. This makes it necessary to modify and fine-tune the framework. In this chapter, we take a more modular approach to deontic STIT logic by adopting *possible world semantics* instead of utility functions (cf. Chapter 2). In particular, we adopt a non-normal modal approach to obligations, using neighborhood semantics (Chellas, 1980). We provide sound and complete axiomatizations for the entire class of deontic STIT logics accommodating the various kinds of OiC principles (Sections 3.2 and 3.3).

Third, we use the resulting deontic STIT logics to obtain a formal taxonomy of the OiC readings discussed. We classify the ten OiC principles according to the respective strength of the underlying STIT logics in which they are embedded (Section 3.4). Furthermore, we determine which logics subsume each other. This gives rise to what we call an *endorsement principle*. Namely, it demonstrates which endorsement of which OiC readings logically commits one to endorse other OiC readings (from the vantage of STIT). The logics are also applied to show the mutual independence of various OiC readings.

Fourth, we compare the variety of OiC with other metaethical principles (Section 3.5). The results are twofold: First, we determine which readings of OiC imply or are logically implied by other metaethical principles. Second, we show under which metaethical principles specific differences between OiC readings disappear. We argue that most metaethical principles are significantly related to OiC. Similar to the endorsement principle, we determine which metaethical principles force one to adopt particular interpretations of OiC and vice versa (in the context of STIT).

Last, to determine the relations between different OiC readings as accurate as possible, we must abstain from adopting other deontic reasoning properties (such as the aggregation of obligations). A common objection to adopting a non-normal modal approach to deontic logic is that certain intuitively desirable inferences are lost, and the logic in question becomes too weak (Van Fraassen, 1973; Horty, 1997). To satisfactorily address this objection, we introduce several extensions of the developed class of deontic STIT logics that reintroduce deontic reasoning principles that simultaneously take into account what 'can be done' (Section 3.5). Among others, we enhance the developed deontic STIT logics with a restricted form of deontic aggregation conjoining only consistent obligations.

Differences. The present chapter is a continuation of the work in (van Berkel and Lyon, 2021). In that work, we used *normal modal operators* to characterize obligations and left the axiomatization of various *deliberative* OiC principles for future work. In the present chapter, we adopt a *non-normal modal* approach which enables the sound and complete axiomatization of all ten formalized OiC principles in (van Berkel and Lyon, 2021). Furthermore, by adopting a non-normal modal approach, we can provide a more fine-grained analysis of OiC. This approach led to the following novel contributions:

- We provide an alternative, more accurate, formalization of the OiC principle 'ought implies logically possible' (Section 3.1).
- We semantically characterize the obligation modality using neighborhood semantics instead of relational semantics (Section 3.2) and show soundness and completeness of the entire class of logics (Section 3.3).
- The logical taxonomy of OiC principles is exhaustive for all ten principles, including the deliberative readings of OiC (Section 3.4).
- We discuss other metaethical principles in the context of OiC and extend the class of logics with several intuitive deontic reasoning principles (Section 3.5).

As a final remark, in (van Berkel and Lyon, 2021), we used (labeled) sequent-style calculi (Negri, 2005) instead of axiomatic systems. The upshot of that approach is that we can potentially use proof-search methods in the context of deontic STIT logic. Although some results were obtained in (Lyon and Berkel, 2019; Lyon, 2021), it is left to future work to develop calculi for the logics of this chapter.

Outline. We analyze ten readings of OiC in Section 3.1. In Section 3.2, we introduce the class of OS logics for the analysis of Ought Implies Can in STIT. We prove soundness and completeness for all the logics of this class in Section 3.3. After that, we analyze the logical taxonomy of OiC in Section 3.4. In Section 3.5, we extend these logics with other metaethical principles and several deontic reasoning principles. Related work and future research are addressed in Section 3.6.

3.1 Ought Implies Can: Ten Interpretations

Disagreement on OiC can be best understood in terms of the degree to which an agent must be burdened with or relieved from duties (Vranas, 2007). Such discussions revolve around the appropriate interpretation of the terms 'ought', 'implies', and predominantly, 'can'. In what follows, we take 'ought' to represent an agent-dependent obligation and take 'implies' to stand for material implication. We refer to the works of van Ackeren and Kühler (2015) and Vranas (2007) for a detailed discussion of the terms 'ought' and 'implies'. In this section, we introduce and discuss ten important interpretations of OiC. We focus on different interpretations of the term 'can' and roughly identify four categories: 'can' as possibility, 'can' as ability, 'can' as violability, and 'can' as control. These four concepts give rise to eight OiC principles. We end this section discussing two additional OiC principles that receive a normative reading of the term 'can'.

Throughout this section, we introduce logical formalizations of the various OiC readings. We employ the (atemporal) deontic STIT language $\mathcal{L}_n^d$ (Definition 2.1, page 29) and refer to Chapter 2 for a detailed discussion of the language. We briefly recall some notation: we let φ stand for an arbitrary formula from $\mathcal{L}_n^d$. The connectives $\neg, \wedge$, and $\rightarrow$ are respectively interpreted as 'not', 'and', and 'implies'. Let $\top$ and $\bot$ denote 'tautology', respectively 'contradiction'. Let $[i]$ be the basic STIT operator expressing "agent i sees to it that" (some proposition holds). Alternatively, we take $[i]\varphi$ to express that "agent i exercises a choice that ensures φ". We use the operator $\Box$ to denote that (some proposition) "is currently settled true". Alternatively, we read a formula $\Box\varphi$ as "φ is realized at the present moment".[2] The main use of $\Box$ is to discern between those states of affairs that are realizable through an agent's choice and those that are realized irrespective of the agents' choices. We take $\Diamond$ as the dual of $\Box$, denoting that some state of affairs is currently realizable.

Last, the deontic modality $\otimes_i$ is read as "it ought to be the case for agent i that". We stress that OiC is essentially agentive but does not necessarily refer to choice in particular. For this reason, we adopt "it ought to be the case for agent i that" instead of the stronger "agent i ought to see to it that". The latter reading corresponds to the *quasi-agentive* reading of obligation, as advocated by Belnap and Perloff (1988) and adopted by Horty (2001). In Section 3.6, we investigate the logical consequences for OiC when adopting the quasi-agentive reading of obligation.

Ought implies Logical Possibility. The first principle, one of the weakest interpretations of OiC, merely requires the content of an agent's obligation to be non-contradictory. It is phrased and formalized as follows:

What is obligatory for an agent, is logically consistent: $\otimes_i\varphi \rightarrow \neg \otimes_i \bot$ (OiLP).

[2] Our focus is on atemporal readings of OiC and, therefore, it suffices to refer to moments as isolated events at which agents exercise choices. See Section 3.2 for formal details and Section 3.6 for a discussion of temporal readings of OiC.

OiLP expresses that if anything is obligatory, then there is no obligation to bring about what is (logically) impossible, i.e., $\neg \otimes_i \bot$. Within the philosophical literature, this interpretation is referred to as “ought implies logical possibility” (Vranas, 2007), and the principle is often equated with the “deontic consistency” principle, e.g., see (van Eck, 1982; Lemmon, 1962).[3] As a minimal constraint on deontic reasoning, the principle is a cornerstone of deontic logic (Anderson and Moore, 1957; Hilpinen and McNamara, 2013; von Wright, 1951). Still, some have repudiated it, e.g., Lemmon (1962). In Section 3.5, we discuss Lemmon’s argument and pinpoint what, we believe, goes awry in his rejection of OiC.

Remark 3.1. OiLP *ensures that each obligatory φ is logically consistent and, consequently, not equivalent to $\bot$. To see this point, let $\otimes_i\varphi$ be an obligation and suppose φ is inconsistent. Then, $\varphi \equiv \bot$ and so we infer $\otimes_i\bot$. This inference is valid in any (non-) normal modal logic by the validity of the rule of congruence (Blackburn et al., 2004; Chellas, 1980). We observe that the formalization of* OiLP *differs from the one given in (van Berkel and Lyon, 2021). There, we formalized* OiLP *as* $(\dagger)\ \otimes_i \varphi \rightarrow \neg \otimes_i \neg\varphi$. *In fact, for the deontic* STIT *logic* DS_n *from Chapter 2 these two formulae are equivalent. The equivalence is due to the normality of the $\otimes_i$ operator. In this chapter, we adopt a non-normal modal approach to the operator $\otimes_i$, which allows us to introduce certain refinements in how we axiomatize OiC. In a non-normal modal logic setting* OiLP *and* $(\dagger)$ *are not equivalent. We come back to this in Section 3.5.*

Ought implies Realizability. The next interpretation refers to what is realizable at a given moment. It is formulated as follows:

What is obligatory for an agent, is realizable: $\otimes_i\varphi \rightarrow \Diamond\varphi$ (OiRz).

This reading of OiC requires that everything which is obligatory is realizable at the moment in which the agent must choose. Consequently, that which is obliged is compatible with some of the agent’s choices. Nevertheless, OiRz remains an agent-independent principle in the following sense: Suppose I am obliged to open the window. Then, OiRz requires that an open window is currently realizable—e.g., it is not jammed—*even though* I cannot open it myself due to being tied to the chair. Arguably, the agent-independent readings of ‘can’ in OiLP and OiRz are too weak to capture the more common philosophical interpretations of OiC.[4] For instance, although a moon eclipse is both logically possible and realizable, it should not be considered something an agent ought to bring about. For this reason, most interpretations of ‘can’ involve the agent explicitly.

[3] We point out that von Wright (1981) calls OiLP ‘Bentham’s Law’ and remarks that Mally already adopted it in what is known as the first attempt to construct a deontic logic (Lokhorst, 1999).

[4] Hilpinen and McNamara (2013) refer to OiC as ‘Kant’s law’ and classify OiLP and OiRz as weak versions of this law. However, it is open to debate which reading of OiC (if any) Kant would endorse, e.g., see (Kohl, 2015; Timmermann, 2013).

Ought implies Ability. The following OiC reading enforces an explicitly agentive precondition on obligations:

What is obligatory for an agent, can be seen to by the agent: $\otimes_i\varphi \rightarrow \Diamond[i]\varphi$ (OiA)

In other words, OiA requires the agent's ability to guarantee (through choice) the realization of that which is prescribed.[5] The concept of ability has many formulations: for example, it may denote general ability, current ability, potential ability, learnability, know-how, and even technical skill.[6] In what follows, we take 'ability' to mean that the agent in question can guarantee a certain outcome by exercising a *choice* at the current moment.

Observe that OiA is the principle implied by Horty's (2001) utilitarian deontic STIT logic (discussed as US_n in Chapter 2). However, this OiC reading does not completely capture the notion of 'ability' as predominantly encountered in the philosophical literature. That is, OiA merely requires that what is prescribed for the agent can be guaranteed through one of the agent's choices but does not exclude *vacuously satisfied obligations*. Namely, agents can still be obliged to bring about inevitable states of affairs. Think of an obligation to realize the tautological state of affairs "the door is open or the door is not open". In the context of obligations, philosophical notions of ability often exclude such consequences by strengthening the concept of ability with either (i) the *possibility* that the obligation may be *violated*, (ii) the agent's *ability to violate* the obligation (i.e., the agent may refrain from fulfilling her duty), (iii) the right *opportunity* for the agent to exercise her ability, or (iv) the agent's *control* over the situation (i.e., the agent's power to decide over the fate of that which is prescribed). In a deontic context, the above four notions ensure that obligations range over states of affairs that are capable of being otherwise. According to Horty and Belnap (1995), the latter is a precondition for *deliberative agency*. For this reason, we refer to the following OiC interpretations—based on (i)-(iv)—as deliberative.

Ought implies Violability. This principle requires the violability of an obligation, which means that the complement of what is prescribed must be realizable.

An agent's obligations are violable: $\otimes_i\varphi \rightarrow \Diamond\neg\varphi$ (OiV)

Governatori and Rotolo (2006) argue that for obligations to be meaningful at all, they must be violable. Namely, since obligations provide reasons to act and tautological obligations are gratuitously observed, the latter do not provide any reasons for behaving in one way

[5] Similarly, von Wright (1968, p.50) distinguishes between human and physical possibility (cf. OiA and OiRz, respectively), both implying logical possibility (cf. OiLP) as a necessary condition.

[6] See the works of Broersen (2011b), Brown (1988), Goldman (1970), and von Wright (1963a) for various notions of 'ability'. We refer to McConnell (1989), Stocker (1971), and Vranas (2007) for discussions on the related notion of 'inability'.

rather than another. Hence, tautological obligations are meaningless. The principle OiV excludes such meaningless obligations. That is, given OiV, a tautological obligation $\otimes_i\top$ would imply the possibility of a contradiction, i.e., $\Diamond\bot$ (where $\bot := \neg\top$). Furthermore, OiV strongly relates to the metaethical principle of "no vacuous commands," which ensures that neither tautologies are obligatory nor contradictions are prohibited (von Wright, 1963a). We discuss the latter in Section 3.5. Just as for OiLP and OiRz, violations are not necessarily agent-dependent: a violation might still arise through causes external to the agent. For instance, the window that ought to be opened by me might be closed through a strong gust of wind.

Ought implies Refrainability. The following principle strengthens the notion of violability by making it an agentive matter:

Obligations are deliberately violable by the agent: $\otimes_i\varphi \rightarrow \Diamond[i]\neg[i]\varphi$ (OiRef).

This OiC reading requires that the agent can refrain from satisfying her obligation. In the jargon of STIT, *refraining* from fulfilling one's duty requires "an embedding of a non-acting within an acting" (Belnap et al., 2001, p.43). That is, it requires the possibility to "see to it that one does not see to it that" (some proposition holds). However, the two principles OiV and OiRef may be insufficient as OiC principles when that which is obliged is not possible in the first place.[7] For instance, it is not difficult for an agent to violate an obligation to create a moon eclipse. We often find the ideas from the previous five OiC interpretations combined to avoid such cases. We discuss three such principles.

Ought implies Opportunity. This principle combines the two interpretations of 'can' as 'realizable' and 'violable'. The result is that obligations range over contingent states of affairs:

What is obligatory for an agent, is contingent in nature: $\otimes_i\varphi \rightarrow (\Diamond\varphi \wedge \Diamond\neg\varphi)$ (OiO).

The two conjuncts in the consequent of OiO constitute what is referred to as the *opportunity* for an agent to actively fulfill her duty; see (Vranas, 2007; von Wright, 1963a). Accordingly, we use the terms 'opportunity' and 'contingency' interchangeably. We point out that OiO does not state that obligations are contingent but only that what is prescribed by the obligation is a contingent state of affairs. Like previous terms, 'opportunity' and 'contingency' have several readings in the literature (Copp, 2017; Dahl,

[7] We conjecture that this is why Vranas (2007) states that OiRef is strictly not an OiC principle. Furthermore, observe that violability and refrainability strongly relate to the metaethical principle of "alternate possibility," which states that an agent is morally culpable if it could have acted otherwise. Due to the involvement of auxiliary concepts such as culpability, we will not further discuss this principle in this chapter. We refer to Copp (2017) and Yaffe (1999) for an introduction.

1974; Vranas, 2007; von Wright, 1951). What these readings have in common is that they refer to the propriety of the circumstances in which the agent must fulfill her duty. At a minimum, both opportunity and contingency require that the prescribed state of affairs is manipulable, i.e., the state of affairs can become true or false.[8] This interpretation of OiO relates to what von Wright (1963a) has in mind when he talks about the opportunity to interfere with the course of nature. In Section 3.5, we provide a detailed discussion of OiO in relation to the principle of "Deontic Contingency".

Ought implies Ability and Opportunity. Furthermore, 'can' may also be taken as the agent's *ability* together with the right *opportunity*. Following Vranas (2007), the latter component specifies the situation hosting the event in which the agent has to exercise her ability. The following principle brings these ideas together:

> *What is obligatory for an agent, is a contingent state of affairs whose truth the agent has the ability to secure*: $\otimes_i \varphi \rightarrow (\Diamond[i]\varphi \wedge \Diamond\varphi \wedge \Diamond\neg\varphi)$ (OiA+O).[9]

The above formulation is the first completely agentive interpretation of OiC, i.e., making that which is obligatory fall, in all its facets, within reach of the agent. Such a reading of OiC can be considered genuinely deliberative, and both Vranas (2007) and von Wright (1963a) appear to endorse a principle similar to OiA+O.

Ought implies Control. Last, we consider an OiC reading which restricts obligations to those states of affairs within the agent's complete *control*.

> *The agent has the ability to see to it that the obligation is fulfilled and has the ability to see to it that the obligation is violated*: $\otimes_i \varphi \rightarrow (\Diamond[i]\varphi \wedge \Diamond[i]\neg\varphi)$ (OiCtrl).

This reading, arguably advocated by Stocker (1971), requires that an agent can act *freely* when under obligation: "it has often been maintained that we act freely in doing or not doing an act only if we both can do it and are able not to do it" (p.305).[10] This instance of OiC implies that an agent is only subject to obligations whose subject matter is within the agent's *power*. The principle is arguably too strong: by restricting obligations to situations in which the agent is in complete control, one excludes those scenarios in

[8] We point out that 'opportunity' and 'contingency' are not synonyms and a more fine-grained distinction is possible. For instance, in temporal settings, a state of affairs can occasionally be true and false (i.e., contingent), even though, at the present moment, the state of affairs is settled true and thus beyond the scope of the agent's influence (i.e., there is no opportunity). We will not explore this refinement in this chapter.

[9] In basic STIT logic, the occurrence of $\Diamond\varphi$ in the consequent of OiA+O can be omitted since it is implied by $\Diamond[i]\varphi$. That is, if φ can be the result of an agent's choice, then by definition, it is realizable. The formula $\Diamond\varphi$ is part of OiA+O for the sake of completion.

[10] In the above quote, "able not to do [φ]" can also be formally interpreted as $\Diamond[i]\neg[i]\varphi$, instead of $\Diamond[i]\neg\varphi$. The resulting principle would be equivalent to OiA+O because $\Diamond[i]\neg[i]\varphi$ is equivalent to $\Diamond\neg\varphi$ in basic STIT logic (Belnap et al., 2001).

which the agent might only partially, but not decisively, contribute to securing an ideal situation. For instance, think of group behavior in which agents need to coordinate.

Normative Readings of Ought Implies Can. OiC has been regarded as too strong to be imposed on ethical theories and normative codes. For example, Lemmon (1962) rejects the legitimacy of OiLP in light of the existence of moral dilemmas (we discuss Lemmon's argument at length in Section 3.5). Others have adopted meta-standpoints toward OiC. For instance, Hintikka (1970) argues that OiC is only dispositional, merely expressing a normative attitude towards the principle.Two options present themselves in this respect: (i) "it *ought to be* that OiC holds" and (ii) "it *ought to be possible* for an agent to fulfill her obligations". Hintikka (1970) seems to advocate the first option, which is intuitively formalized as $\mathcal{O}(\otimes_i \varphi \rightarrow \Diamond\varphi)$ (where the first obligation is an agent-independent 'ought to be' modality). However, option (i) is not an OiC principle. It only expresses that OiC *should* hold as a metaethical principle; cf. (McConnell, 1985). The second option (ii) is indeed an OiC principle, and we consider two possible interpretations.

The first one we refer to as *Ought implies Normatively Can* and is phrased accordingly:

> *What is obligatory for an agent, ought to be realizable*: $\otimes_i \varphi \rightarrow \otimes_i \Diamond\varphi$ (OiNC).

The second interpretation adopts an agent-dependent reading of 'can'. We call this principle *Ought implies Normatively Able*:

> *What is obligatory for an agent, ought to be realizable through the agent's choice*: $\otimes_i \varphi \rightarrow \otimes_i \Diamond[i]\varphi$ (OiNA).

We point out that we interpret the obligation in the consequent of OiNC and OiNA as *agent-dependent*. Thus, OiNC reads "If φ is obligatory for agent i, then it ought to be the case for agent i that φ is realizable". Since obligations in the antecedent of these principles are agent-dependent, we consider it more accurate to say that what is obligatory ought to be realizable *for that agent*, thus making explicit reference to the agent in question. Normative interpretations of the first eight OiC interpretations are straightforwardly obtained. For the aims of this chapter, the principles OiNC and OiNA suffice.

In Table 3.1, the ten principles are collected and associated with references to the various authors discussing them. We stress that the references in Table 3.1 relate to the works that (philosophically) discuss ideas about these principles. The corresponding formalizations are our own and may not correspond with those given in the references (if any are given). Furthermore, the list of discussed principles is not exhaustive, and in Section 3.6, we briefly discuss some alternative OiC interpretations from the literature. It is not our aim to decide which OiC principle should be adopted, and notable cases were for each of them. Instead, we aim to provide a logical investigation of how these principles relate. To this, we turn now.

Name	Ought implies...	Formalized	Literature
OiLP	Logical Possibility	$\otimes_i \varphi \to \neg \otimes_i \bot$	Anderson and Moore (1957), van Eck (1982), and von Wright (1951; 1981)
OiRz	Realizability	$\otimes_i \varphi \to \Diamond \varphi$	van Eck (1982), Hilpinen and McNamara (2013), and Horty (2001, Ch.3)
OiA	Ability	$\otimes_i \varphi \to \Diamond [i] \varphi$	Horty (2001, Ch.4) and von Wright (1963a, Ch.7)
OiV	Violability	$\otimes_i \varphi \to \Diamond \neg \varphi$	Anderson and Moore (1957), Dahl (1974), Goldman (1970), and von Wright (1963a, Ch.8)
OiRef	Refrainability	$\otimes_i \varphi \to \Diamond [i] \neg [i] \varphi$	Goldman (1970) and Vranas (2018a)
OiO	Opportunity	$\otimes_i \varphi \to (\Diamond \varphi \land \Diamond \neg \varphi)$	Anderson and Moore (1957), Copp (2017), Dahl (1974), and von Wright (1951; 1968)
OiA+O	Ability and Opp.	$\otimes_i \varphi \to (\Diamond [i] \varphi \land \Diamond \varphi \land \Diamond \neg \varphi)$	van Ackeren and Kühler (2015), Kohl (2015), Vranas (2007), and von Wright (1963a)
OiCtrl	Control	$\otimes_i \varphi \to (\Diamond [i] \varphi \land \Diamond [i] \neg \varphi)$	Dahl (1974), Stocker (1971), and McConnell (1989)
OiNC	Normatively Can	$\otimes_i \varphi \to \otimes_i \Diamond \varphi$	van Ackeren and Kühler (2015) and Hintikka (1970)
OiNA	Normatively Able	$\otimes_i \varphi \to \otimes_i \Diamond [i] \varphi$	van Ackeren and Kühler (2015) and Hintikka (1970)

Table 3.1: List of the ten OiC principles and their treatment in the literature.

3.2 Deontic STIT Logics: a Non-Normal Modal Approach

In this section, we introduce deontic STIT logics for each reading of OiC. To differentiate the resulting logics from the logics developed in Chapter 2, we write OS_n to indicate that the logic serves the analysis of Ought-implies-Can in STIT (with n referring to the number of agents). In (van Berkel and Lyon, 2021), we formalized the *deliberative* OiC principles—OiV, OiRef, OiO, OiA+O, and OiCtrl—by means of *defined* modal operators. This section demonstrates that all OiC principles can be axiomatized and semantically characterized once we move to a non-normal modal interpretation of the deontic modality $\otimes_i$. This is Objective 1.

Definition 3.1 (The Language $\mathcal{L}^d_n$). *Let* $\mathsf{Atoms} = \{p, q, r, \dots\}$ *be a denumerable set of propositional atoms and let* $\mathsf{Agents} = \{1, 2, \dots, n\}$ *be a finite set of agent labels. The* *language $\mathcal{L}_n^d$ is defined via the following BNF grammar:*

$$\varphi ::= p \mid \neg\varphi \mid \varphi \wedge \varphi \mid \Box\varphi \mid [i]\varphi \mid \ \otimes_i \varphi$$

where $p \in \mathsf{Atoms}$ and $i \in \mathsf{Agents}$.

Let $\Diamond$, $\langle i \rangle$, and $\ominus_i$ be the duals of $\Box$, $[i]$, and $\otimes_i$, respectively. For a discussion of the language $\mathcal{L}_n^d$ we refer to Section 2.1. Whether $\otimes_i$ captures the quasi-agentive reading of obligation depends on whether the formula $\otimes_i\varphi \equiv \otimes_i[i]\varphi$ is valid in the logic in question (see Section 2.1, Remark 2.1). In order to capture certain nuances of OiC, we forego the quasi-agentive reading and interpret $\otimes_i$ as "it is obligatory for agent i that" (some proposition holds). In Section 3.6, we discuss the logical consequences of adopting the quasi-agentive reading of $\otimes_i$ for the analysis of OiC.

3.2.1 Axiomatizations of OiC in Deontic STIT

The Hilbert-style axiomatization of the minimal logic OS_n is given below.

Definition 3.2 (The Axiomatization of OS_n). *We define OS_n to be the following collection of axiom schemes and rules:*

A0. *All classical propositional tautologies;*

R0. *From φ and $\varphi \to \psi$, infer ψ;*

A1. $\Box(\varphi \to \psi) \to (\Box\varphi \to \Box\psi)$;

A2. $\Box\varphi \to \varphi$;

A3. $\Diamond\varphi \to \Box\Diamond\varphi$;

A4. $[i](\varphi \to \psi) \to ([i]\varphi \to [i]\psi)$;

A5. $[i]\varphi \to \varphi$;

A6. $\langle i \rangle\varphi \to [i]\langle i \rangle\varphi$;

A7. $\Box\varphi \to [i]\varphi$;

A8. $\bigwedge_{i\in\mathsf{Agents}} \Diamond[i]\varphi_i \to \Diamond(\bigwedge_{i\in\mathsf{Agents}}[i]\varphi_i)$;

A10. $\otimes_i\varphi \to \Box \otimes_i \varphi$;

R1. *From φ, infer $\Box\varphi$;*

R2. *From $\varphi \equiv \psi$, infer $\otimes_i\varphi \equiv \otimes_i\psi$;*

where we have a copy of A4–A7*,* A10*, and* R2 *for each $i \in$* Agents*. The* logic OS_n *is the smallest set of formulae from $\mathcal{L}_n^d$ closed under all instances of the axiom schemes and applications of the inference rules* R0 *–* R2*. Whenever $\varphi \in \mathsf{OS}_n$, we say that $\varphi \in \mathcal{L}_n^d$ is a* OS_n-theorem *and write $\vdash_{\mathsf{OS}_n} \varphi$. Last, OS_n-derivability is defined as usual (see Definition 2.3).*

Axioms A1-A7 and R1, characterize $\Box$ and $[i]$ as normal modal S5-operators. In particular, necessitation holds for $[i]$ by virtue of R1 and A7. The bridge axiom A7 confines choices to moments. Axiom A8 denotes the independence of agents property of STIT. For a discussion of these non-deontic axioms of OS_n see Section 2.1. There is one deontic axiom A10 expressing the fact that obligations are moment dependent. Namely, obligations express which continuations of the present moment are deontically ideal for that agent. For that reason, obligations do not vary from world to world within a moment but hold independently of any of the agent's choices at that moment. Notice that $\otimes_i$ is a *non-normal* modal operator, i.e., we neither adopt the distribution axiom A9 $\otimes_i(\varphi \rightarrow \psi) \rightarrow (\otimes_i\varphi \rightarrow \otimes_i\psi)$ of Definition 2.2 nor a rule of necessitation stating that $\vdash_{\mathsf{OS}_n} \varphi$ implies $\vdash_{\mathsf{OS}_n} \otimes_i\varphi$ (hence, the missing A9 in Definition 3.2). We only adopt the rule R2, which captures the minimal property of non-normal modal logics, referred to as the rule of congruence (Chellas, 1980). Intuitively, the rule enables us to substitute equivalent formulae inside the scope of a deontic modality $\otimes_i$. It ensures, for instance, that redundant syntactic differences do not influence the derivability of $\otimes_i$-formulae, e.g., think of $\varphi \equiv (\varphi \wedge \varphi)$. Variations of R2 hold for $\Box$ and $[i]$ due to the fact that these modalities are normal. Namely, suppose $\vdash_{\mathsf{OS}_n} \varphi \equiv \psi$, by R1 we obtain $\vdash_{\mathsf{OS}_n} \Box(\varphi \equiv \psi)$ and by modal reasoning using A1 and R0 we derive $\vdash_{\mathsf{OS}_n} \Box\varphi \equiv \Box\psi$.

Remark 3.2. *We point out in passing that OS_n does not satisfy the bridge axiom $\Box\varphi \rightarrow \otimes_i\varphi$. A consequence of adopting that axiom is the theorem $\otimes_i\top$, which conflicts with any deliberative reading of OiC (cf. the discussion on page 88). Furthermore, we point out that Horty's (2001) Utilitarian Deontic* STIT*—i.e., US_n of Section 2.3—does contain the above bridge axiom. Here, we omit this axiom to capture some of the nuances we find in the literature of OiC. For instance, due to the bridge axiom $(\otimes_i\varphi \rightarrow \neg \otimes_i \bot) \equiv (\otimes_i\varphi \rightarrow \Diamond\varphi)$ is a theorem of US_n, which means that the logic cannot differentiate between* OiLP *and* OiRz*.*

The axiomatic system OS_n may be extended with any (combination) of the ten formalized OiC principles from Table 3.1. The resulting logics are defined in Definition 3.3.

Definition 3.3 (The Logic $\mathsf{OS}_n\mathsf{X}$). *The logic $\mathsf{OS}_n\mathsf{X}$ is defined as the extension of OS_n with the axiom schemes in $\mathsf{X} \subseteq \{\mathsf{A}i \mid 11 \leq i \leq 20\}$, where* A11,...,A20 *are the following axiom schemes:*

A11. $\otimes_i\varphi \rightarrow \neg \otimes_i \bot$ (OiLP)*;*

A12. $\otimes_i\varphi \rightarrow \Diamond\varphi$ (OiRz)*;*

A13. $\otimes_i\varphi \to \Diamond[i]\varphi$ (OiA);

A14. $\otimes_i\varphi \to \Diamond\neg\varphi$ (OiV);

A15. $\otimes_i\varphi \to \Diamond[i]\neg[i]\varphi$ (OiRef);

A16. $\otimes_i\varphi \to (\Diamond\varphi \wedge \Diamond\neg\varphi)$ (OiO);

A17. $\otimes_i\varphi \to (\Diamond[i]\varphi \wedge \Diamond\varphi \wedge \Diamond\neg\varphi)$ (OiA+O);

A18. $\otimes_i\varphi \to (\Diamond[i]\varphi \wedge \Diamond[i]\neg\varphi)$ (OiCtrl);

A19. $\otimes_i\varphi \to \otimes_i\Diamond\varphi$ (OiNC);

A20. $\otimes_i\varphi \to \otimes_i\Diamond[i]\varphi$ (OiNA);

for each $i \in$ Agents. The inference relation $\vdash_{\mathsf{OS}_n\mathsf{X}}$ is defined as usual.

Definition 3.3 yields arbitrary extensions of the minimal logic OS_n. We make three points: First, we are mostly interested in logics $\mathsf{OS}_n\mathsf{X}$ where X contains only a single OiC axiom scheme for each agent. In other words, these are deontic STIT logics tailored to particular readings of OiC. Second, on a related note, one could adopt for different agents different OiC principles. Although the formal results of this chapter hold for all these logics, we mainly focus on logics in which the same OiC principle applies to every agent. Third, in Section 3.4, we demonstrate that some OiC principles logically imply others. This means that not all $\mathsf{OS}_n\mathsf{X}$ axiomatizations are minimal axiomatizations. To give an example, OiO logically implies OiV and for that reason the logic $\mathsf{OS}_n\mathsf{X}$ with $\mathsf{X} = \{$ A16,A14 $\mid$ for each $i \in$ Agents$\}$ is equivalent to the logic $\mathsf{OS}_n\mathsf{X}'$ with $\mathsf{X}' = \{$ A16 $\mid$ for each $i \in$ Agents$\}$.

Reasons for Using Non-Normal Modalities. Before providing the semantics for $\mathsf{OS}_n\mathsf{X}$, it must be noted that the logics are minimal. Namely, the logics are tailored to axiomatize OiC principles, and so far, no additional properties have been enforced. In particular, the logics do not satisfy the following properties:

M. $\otimes_i(\varphi \wedge \psi) \to (\otimes_i\varphi \wedge \otimes_i\psi)$;

C. $(\otimes_i\varphi \wedge \otimes_i\psi) \to \otimes_i(\varphi \wedge \psi)$;

N. $\otimes_i\top$.

The axioms M, C, and N represent *monotonicity*, *aggregation*, and *necessity*, respectively. They are theorems of any *normal* modal characterization of $\otimes_i$ (Chellas, 1980).[11]

[11] The common name for Axiom M is monotonicity, e.g., see Chellas (1980). Despite its shared name, this modal property must be distinguished from the use of monotonicity as a property of the inference relation $\vdash$, e.g., see Arieli et al. (2021). The latter property is central to part III of this thesis, e.g., see Definition 7.2 on page 258. Axiom N corresponds to the rule of necessitation in normal modal logic.

In (van Berkel and Lyon, 2021), we adopted a normal modal approach to OiC. There, we proceeded in two ways: first, we *defined* deontic STIT operators capturing deliberative aspects of obligation, and second, we introduced a class of axioms determining the behavior of the $\otimes_i$ operator. The two together were sufficient to obtain some first results about the logical taxonomy of OiC. However, as already noted in (van Berkel and Lyon, 2021), the use of defined deliberative deontic operators was *ad hoc*. There, it was left to future work to provide a proper axiomatization of these deliberative OiC principles. We explain this further: A normal modal interpretation of $\otimes_i$ implies that $\otimes_i\top$ is a theorem of the logic. Consequently, an axiomatization of, for instance, 'ought implies violability' (OiV)—i.e., $\otimes_i\varphi \rightarrow \Diamond\neg\varphi$—would render the logic inconsistent. Namely, by the axiom N we can derive $\otimes_i\top$ which together with OiV implies $\Diamond\bot$. The latter is inconsistent with the fact that $\Box$ is also a normal modal operator, i.e., $\Box\neg\bot$ is a theorem. In (van Berkel and Lyon, 2021), this problem was addressed by introducing the *deliberative* obligation,

$$\otimes_i^d\varphi := \otimes_i\varphi \wedge \Diamond\neg\varphi$$

expressing that an agent's obligations can be violated. See (van Berkel and Lyon, 2021) for a discussion of the other defined deliberative obligation. By adopting a non-normal modal approach to $\otimes_i$ in this chapter, we forego these ad hoc definitions and take the ten OiC readings as axioms proper.

A straightforward objection to the approach in this chapter is that the resulting logics are too weak for intuitive deontic reasoning tasks (cf. the absence of M, C, and N). Still, the approach taken here is deliberate: by adopting minimal axiomatizations of $\mathsf{OS}_n\mathsf{X}$, we can better understand how various OiC axioms are logically related. We exclude the risk that certain OiC axioms seem related due to the presence of additional deontic reasoning principles. In order to restore some inferential power in $\mathsf{OS}_n\mathsf{X}$, we extend the logics with restricted versions of M and C that take into account the different OiC axioms in question. We do this in Section 3.5.

3.2.2 Semantics for OiC in Deontic STIT

We adopt relational semantics (Balbiani et al., 2008) to characterize the non-deontic fragment of OS_n and adopt *neighborhood semantics* to capture the various non-normal readings of OiC. Neighborhood frames (Chellas, 1980) were developed to characterize logics that do not satisfy the properties induced by minimal relational frames, i.e., M, C, and N. Instead of adopting a directed relation $\mathcal{R}_{\otimes_i}$ over worlds, we adopt a neighborhood function $\mathcal{N}_{\otimes_i}$ that maps worlds $w, v, u, \dots$ to *sets* of worlds $X, Y, Z \subseteq W$. We say that at w, the set of worlds Z is considered deontically ideal for agent i whenever $Z \in \mathcal{N}_{\otimes_i}(w)$. In what follows, we define minimal OS_n-frames and subsequently provide a list of properties with which these frames may be extended.

Definition 3.4 (Frames and Models for OS_n)**.** *An* OS_n-frame *is defined as a tuple* $\mathfrak{F} = \langle W, \mathcal{R}_\Box, \{\mathcal{R}_{[i]} \mid i \in \mathsf{Agents}\}, \{\mathcal{N}_{\otimes_i} \mid i \in \mathsf{Agents}\}\rangle$*. Let* $\mathcal{R}_\alpha \subseteq W \times W$ *and* $\mathcal{R}_\alpha(w) := \{v \in W \mid (w, v) \in \mathcal{R}_\alpha\}$ *for* $\alpha \in \{\Box\} \cup \{[i] \mid i \in \mathsf{Agents}\}$*. Let* $\mathcal{N}_{\otimes_i}$ *be a* neighborhood function *for each* $i \in \mathsf{Agents}$ *such that* $\mathcal{N}_{\otimes_i} : W \mapsto \wp(W)$*. Let* W *be a non-empty set of worlds* $w, v, u, \dots$ *such that the following hold:*

C1 $\mathcal{R}_{\Box}$ *is an equivalence relation;*

C2 *For all* $i \in \mathsf{Agents}$*,* $\mathcal{R}_{[i]}$ *is an equivalence relation;*

C3 *For all* $i \in \mathsf{Agents}$*,* $\mathcal{R}_{[i]} \subseteq \mathcal{R}_{\Box}$*;*

C4 *For all* $w \in W$ *and all* $u_1, \dots, u_n \in \mathcal{R}_{\Box}(w)$*,* $\bigcap_{i \in \mathsf{Agents}} \mathcal{R}_{[i]}(u_i) \neq \emptyset$*;*

O1 *For all* $w, v \in W$*, and all* $Z \subseteq W$*, if* $Z \in \mathcal{N}_{\otimes_i}(w)$ *and* $v \in \mathcal{R}_{\Box}(w)$*, then* $Z \in \mathcal{N}_{\otimes_i}(v)$*.*

An OS_n-model *is a tuple* $\mathfrak{M} = \langle \mathfrak{F}, V \rangle$ *where* $\mathfrak{F}$ *is an* OS_n*-frame and* V *is a valuation function mapping propositional atoms to subsets of* W*, i.e.,* $V\colon \mathsf{Atoms} \mapsto \wp(W)$*.*

In Definition 3.4, property **C1** stipulates that $R_{\Box}$ are moments. For each agent in the language, **C2** and **C3** partition moments into choices. **C4** imposes the independence of agents principle. For a discussion of the non-deontic STIT properties **C1**–**C4** we refer to Section 2.1.

The only deontic property imposed on minimal OS_n-frames is **O1**, which captures the idea that what is obligatory is settled true for each moment, irrespective of the choices the agents will make at that moment (cf. axiom $\mathsf{A10}$). The property **O1** corresponds to **D1** of DS_n-frames (Section 2.1). We emphasize that the class of OS_n-frames does not require that worlds ideal at a certain moment are realizable at that very moment (Remark 3.2). This means that what is ideal might not be realizable by any of the agents' (combined) choices and might therefore be beyond the grasp of agency. Last, although an OS_n-frame may contain several moments, we abstain from a temporal extension of OS_n. In Section 3.6, we discuss some temporal OiC principles.

The semantic interpretation of $\mathcal{L}_n^d$ is defined as usual. The modality $\otimes_i$ is evaluated with respect to its corresponding neighborhood function.

Definition 3.5 (Semantics of OS_n-models)**.** *Let* $\mathfrak{M}$ *be an* OS_n*-model and let* $w \in W$ *of* $\mathfrak{M}$*. Let* $\|\varphi\|_{\mathfrak{M}} = \{w \in W \mid \mathfrak{M}, w \models \varphi\}$ *be the* truth set *of worlds satisfying* φ *(we often omit the subscript* $\mathfrak{M}$*). The* satisfaction *of a formula* $\varphi \in \mathcal{L}_n^d$ *in* $\mathfrak{M}$ *at* w *is defined accordingly:*

1. $\mathfrak{M}, w \models p$ iff $\mathfrak{M}, w \in V(p)$*;*

2. $\mathfrak{M}, w \models \neg\varphi$ iff *not* $\mathfrak{M}, w \models \varphi$*;*

3. $\mathfrak{M}, w \models \varphi \land \psi$ iff $\mathfrak{M}, w \models \varphi$ *and* $\mathfrak{M}, w \models \psi$*;*

4. $\mathfrak{M}, w \models \Box\varphi$ iff *for all* $v \in \mathcal{R}_{\Box}(w)$*,* $\mathfrak{M}, v \models \varphi$*;*

5. $\mathfrak{M}, w \models [i]\varphi$ iff *for all* $v \in \mathcal{R}_{[i]}(w)$*,* $\mathfrak{M}, v \models \varphi$*;*

6. $\mathfrak{M}, w \models \otimes_i\varphi$ iff $\|\varphi\|_{\mathfrak{M}} \in \mathcal{N}_{\otimes_i}(w)$*.*

Global truth, validity, and semantic entailment are defined as usual, e.g., see Definition 2.5.

The following list of properties enables us to semantically characterize the ten proposed OiC readings of Table 3.1 (Section 3.1). We write **Oi** with $i \in \mathbb{N}$ to denote deontic frame properties required for OiC.

O2 For all $w \in W, Z \subseteq W$, if $Z \in \mathcal{N}_{\otimes_i}(w)$, then $\emptyset \notin \mathcal{N}_{\otimes_i}(w)$;

O3 For all $w \in W, Z \subseteq W$, if $Z \in \mathcal{N}_{\otimes_i}(w)$, then $\mathcal{R}_{\Box}(w) \cap Z \neq \emptyset$;

O4 For all $w \in W, Z \subseteq W$, if $Z \in \mathcal{N}_{\otimes_i}(w)$, there is a $v \in \mathcal{R}_{\Box}(w)$ such that $\mathcal{R}_{[i]}(v) \subseteq Z$;

O5 For all $w \in W, Z \subseteq W$, if $Z \in \mathcal{N}_{\otimes_i}(w)$, then $\mathcal{R}_{\Box}(w) \cap \overline{Z} \neq \emptyset$;

O6 For all $w \in W, Z \subseteq W$, if $Z \in \mathcal{N}_{\otimes_i}(w)$, then $\mathcal{R}_{\Box}(w) \cap Z \neq \emptyset$ and $\mathcal{R}_{\Box}(w) \cap \overline{Z} \neq \emptyset$;

O7 For all $w \in W, Z \subseteq W$, if $Z \in \mathcal{N}_{\otimes_i}(w)$, then there is a $v \in \mathcal{R}_{\Box}(w)$ such that $\mathcal{R}_{[i]}(v) \subseteq Z$, and $\mathcal{R}_{\Box}(w) \cap Z \neq \emptyset$, and $\mathcal{R}_{\Box}(w) \cap \overline{Z} \neq \emptyset$;

O8 For all $w \in W, Z \subseteq W$, if $Z \in \mathcal{N}_{\otimes_i}(w)$, then there are $v, u \in \mathcal{R}_{\Box}(w)$ such that $\mathcal{R}_{[i]}(v) \subseteq Z$ and $\mathcal{R}_{[i]}(u) \subseteq \overline{Z}$;

O9 For all $w \in W, Z \subseteq W$, if $Z \in \mathcal{N}_{\otimes_i}(w)$, then $\{v \in W \mid \mathcal{R}_{\Box}(v) \cap Z \neq \emptyset\} \in \mathcal{N}_{\otimes_i}(w)$;

O10 For all $w \in W, Z \subseteq W$, if $Z \in \mathcal{N}_{\otimes_i}(w)$, then $\{v \in W \mid$ there is a $u \in \mathcal{R}_{\Box}(v)$ such that $\mathcal{R}_{[i]}(u) \subseteq Z\} \in \mathcal{N}_{\otimes_i}(w)$.

As will be shown in Section 3.3, the properties **O2**–**O10** semantically characterize the ten OiC principles of Table 3.1. To illustrate, consider property **O6**. Let $Z = \|\varphi\|$, then the condition ensures that if $\|\varphi\| \in \mathcal{N}_{\otimes_i}(w)$ (i.e., φ is obligatory for agent i from the perspective w), then $\mathcal{R}_{\Box}(w) \cap \|\varphi\| \neq \emptyset$ and $\mathcal{R}_{\Box}(w) \cap \overline{\|\varphi\|} \neq \emptyset$—the latter which is equivalent to $\mathcal{R}_{\Box}(w) \cap \|\neg\varphi\| \neq \emptyset$—consequently, both φ and $\neg\varphi$ are realizable at moment $\mathcal{R}_{\Box}(w)$. In other words, **O6** characterizes the frame property for 'ought implies opportunity' (OiO). The other properties are read similarly, except for **O9** and **O10**. The latter two concern normative readings of OiC. For instance, recall that OiNA expresses the idea that "ought implies ought to be able". Property **O10** captures this idea. Let $Z = \|\varphi\|$, if $\|\varphi\| \in \mathcal{N}_{\otimes_i}(w)$, then the truth-set $\{v \in W \mid$ there is a $u \in \mathcal{R}_{\Box}(v)$ such that $\mathcal{R}_{[i]}(u) \subseteq \|\varphi\|\}$ is in the neighborhood $\mathcal{N}_{\otimes_i}$ of w. Comparing semantic definitions, this truth set is equal to the set $\|\Diamond[i]\varphi\| \in \mathcal{N}_{\otimes_i}(w)$, which means that $\Diamond[i]\varphi$ is obligatory for agent i at w.

$\mathsf{OS}_n\mathsf{X}$ contains:	A11	A12	A13	A15	A14	A16	A17	A18	A19	A20
$\mathsf{OS}_n\mathsf{X}$-frames satisfy:	**O2**	**O3**	**O4**	**O5**	**O5**	**O6**	**O7**	**O8**	**O9**	**O10**

Table 3.2: Correspondence between the construction of $\mathsf{OS}_n\mathsf{X}$ logics and $\mathsf{OS}_n\mathsf{X}$-frames. For instance, if A11 is an axiom of $\mathsf{OS}_n\mathsf{X}$, then we assume that the corresponding class of $\mathsf{OS}_n\mathsf{X}$-frames contains the property **O2**.

Remark 3.3. *We did not include a frame property for the principle* OiRef*. The reason is that* OiV *and* OiRef *are equivalent for any deontic extension of the basic (atemporal)* STIT *logic. This is due to the equivalence* $\Diamond\neg\varphi \equiv \Diamond[i]\neg[i]\varphi$ *which is a valid formula in the context of the underlying non-deontic* STIT *logic. Consequently, condition **O5**, characterizing* OiV*, likewise captures* OiRef*. An alternative frame property for* OiRef *would be: **O5'** For all* $w \in W$, $Z \subseteq W$*, if* $Z \in \mathcal{N}_{\otimes_i}(w)$*, then there is a* $v \in \mathcal{R}_{\Box}(w)$ *such that for all* $u \in \mathcal{R}_{[i]}(v)$ *we have* $\mathcal{R}_{[i]}(u) \cap \overline{Z} \neq \emptyset$*. It can be straightforwardly checked that the consequent of **O5'** is equivalent to that of **O5**. Soundness and completeness in Section 3.3 demonstrate that **O5** suffices for both.*

The above list of frame properties provides a modular way to obtain various extensions of OS_n-frames. We define the entire class of $\mathsf{OS}_n\mathsf{X}$-frames as follows:

Definition 3.6 (Frames and Models for $\mathsf{OS}_n\mathsf{X}$)**.** *An* $\mathsf{OS}_n\mathsf{X}$-frame *is a tuple* $\mathfrak{F} = \langle W, \mathcal{R}_{\Box}, \{\mathcal{R}_{[i]} \mid i \in \mathsf{Agents}\}, \{\mathcal{N}_{\otimes_i} \mid i \in \mathsf{Agents}\}\rangle$ *such that* $\mathfrak{F}$ *satisfies all properties of an* OS_n*-frame (Definition 3.4) expanded with the frame properties that correspond to the axioms in* X *as stipulated in Table 3.2. An* $\mathsf{OS}_n\mathsf{X}$-model *is a tuple* $\langle \mathfrak{F}, V\rangle$ *where* $\mathfrak{F}$ *is an* $\mathsf{OS}_n\mathsf{X}$-frame *and* V *is a valuation function as in Definition 3.4.*

3.3 Soundness and Completeness

Soundness of an $\mathsf{OS}_n\mathsf{X}$ logic is proven in the usual way (Blackburn et al., 2004; Chellas, 1980) (cf. Section 2.2). Due to the modularity of our approach, it suffices to give a single proof for the entire class of $\mathsf{OS}_n\mathsf{X}$ logics.

In this section, we make (often implicit) use of the following lemma.

Lemma 3.1. *Let* $\mathfrak{M}$ *be an* $\mathsf{OS}_n\mathsf{X}$*-model from Definition 3.6. For each* $w, v \in W$, $[\alpha] \in \{\Box\} \cup \{[i] \mid i \in \mathsf{Agents}\}$*, and* $\varphi, \psi \in \mathcal{L}_n^d$ *we have:*

1. $v \in \mathcal{R}_{[\alpha]}(w)$ *iff* $\mathcal{R}_{[\alpha]}(w) = \mathcal{R}_{[\alpha]}(v)$*;*
2. $\mathcal{R}_{[\alpha]}(w) \cap \|\varphi\| \neq \emptyset$ *iff for all* $v \in \mathcal{R}_{[\alpha]}(w), \mathfrak{M}, v \models \langle\alpha\rangle\varphi$*;*
3. $\mathcal{R}_{[\alpha]}(w) \subseteq \|\varphi\|$ *iff* $\mathfrak{M}, w \models [\alpha]\varphi$*;*
4. $\mathcal{R}_{[\alpha]}(w) \neq \emptyset$*;*

5. $\overline{\|\varphi\|} = \|\neg\varphi\|$ *and* $\|\varphi \wedge \psi\| = \|\varphi\| \cap \|\psi\|$.

Proof. The proofs of (1)-(5) are straightforward by the fact that $\mathcal{R}_\Box$ and $\mathcal{R}_{[i]}$ are equivalence classes, the semantic definitions of $\Box$, $[i]$, $\neg$, and $\wedge$. QED

Theorem 3.1 (Soundness of $\mathsf{OS}_n\mathsf{X}$)**.** *Let* $\mathsf{OS}_n\mathsf{X}$ *be a logic from Definition 3.3. For any formula* $\varphi \in \mathcal{L}_n^d$*, and any* $\Gamma \subseteq \mathcal{L}_n^d$*: if* $\Gamma \vdash_{\mathsf{OS}_n\mathsf{X}} \varphi$*, then* $\Gamma \models_{\mathsf{OS}_n\mathsf{X}} \varphi$*.*

Proof. It suffices to demonstrate the following claim:

$$(\dagger) \text{ if } \vdash_{\mathsf{OS}_n\mathsf{X}} \varphi, \text{ then } \models_{\mathsf{OS}_n\mathsf{X}} \varphi.$$

We prove (†) by demonstrating that all axioms are $\mathsf{OS}_n\mathsf{X}$-valid, and the logical rules of $\mathsf{OS}_n\mathsf{X}$ preserve truth on the respective frame classes. We observe that the non-deontic frame properties **C1**–**C4** of $\mathsf{OS}_n\mathsf{X}$-frames are the same as for the logic DS_n (Chapter 2). Validity of the corresponding axioms is shown in the same way as for DS_n. These cases are therefore omitted (cf. Theorem 2.1). The same reasoning applies to the rules R0 and R1. We show the validity of the deontic axioms and rule R2.

Take an arbitrary $\mathsf{OS}_n\mathsf{X}$-model $\mathfrak{M}$ and an arbitrary $w \in W$ of $\mathfrak{M}$. Note that we assume X to correspond to the frame properties of $\mathfrak{M}$ for which we prove validity, as stipulated in Table 3.2. For example, in proving the validity of A11 $\in$ X we assume $\mathfrak{M}$ satisfies **O2**. In what follows, we omit reference to $\mathfrak{M}$.

A10 $\otimes_i\varphi \rightarrow \Box \otimes_i \varphi$. Assume $\mathfrak{M}, w \models \otimes_i\varphi$ and consider an arbitrary $v \in \mathcal{R}_\Box(w)$. By the semantic definition of $\otimes_i$ we know $\|\varphi\| \in \mathcal{N}_{\otimes_i}(w)$ and so, by property **O1**, we know $\|\varphi\| \in \mathcal{N}_{\otimes_i}(v)$. Since v is arbitrary, we know by the semantic definition of $\Box$ that $Mo, w \models \Box \otimes_i \varphi$.

A11 $\otimes_i\varphi \rightarrow \neg \otimes_i \bot$. Assume $Mo, w \models \otimes_i\varphi$. By the semantic definition of $\otimes_i$ we know $\|\varphi\| \in \mathcal{N}_{\otimes_i}(w)$ and so, by property **O2**, we know that $\emptyset \notin \mathcal{N}_{\otimes_i}(w)$. Since $\emptyset = \|\bot\|$ we know that $\mathfrak{M}, w \models \neg \otimes_i \bot$.

A12 $\otimes_i\varphi \rightarrow \Diamond\varphi$. Assume $\mathfrak{M}, w \models \otimes_i\varphi$. By the semantic definition of $\otimes_i$ we know $\|\varphi\| \in \mathcal{N}_{\otimes_i}(w)$ and so, by property **O3**, we know that $\mathcal{R}_\Box(w) \cap \|\varphi\| \neq \emptyset$. Hence, by Lemma 3.1-(2) we have $\mathfrak{M}, w \models \Diamond\varphi$.

A13 $\otimes_i\varphi \rightarrow \Diamond[i]\varphi$. Assume $\mathfrak{M}, w \models \otimes_i\varphi$. By the semantic definition of $\otimes_i$ we know $\|\varphi\| \in \mathcal{N}_{\otimes_i}(w)$ and so, by property **O4**, we know there is a $v \in \mathcal{R}_\Box(w)$ such that $\mathcal{R}_{[i]}(v) \subseteq \|\varphi\|$. By Lemma 3.1-(3) it follows that $\mathfrak{M}, v \models [i]\varphi$ and therefore $\mathfrak{M}, w \models \Diamond[i]\varphi$.

A14 $\otimes_i\varphi \rightarrow \Diamond\neg\varphi$. Assume $\mathfrak{M}, w \models \otimes_i\varphi$. By the semantic definition of $\otimes_i$ we know $\|\varphi\| \in \mathcal{N}_{\otimes_i}(w)$. By property **O5**, we know that $\mathcal{R}_\Box(w) \cap \overline{\|\varphi\|} \neq \emptyset$ and by Lemma 3.1-(5) we have $\mathcal{R}_\Box(w) \cap \|\neg\varphi\| \neq \emptyset$. So, by Lemma 3.1-(2), $\mathfrak{M}, w \models \Diamond\neg\varphi$.

A15 $\otimes_i\varphi \to \Diamond[i]\neg[i]\varphi$. Assume $\mathfrak{M}, w \models \otimes_i\varphi$. By the semantic definition of $\otimes_i$ we know $\|\varphi\| \in \mathcal{N}_{\otimes_i}(w)$. By property **O5** and Lemma 3.1-(5), we know that $\mathcal{R}_{\Box}(w) \cap \|\neg\varphi\| \neq \emptyset$. Without loss of generality, let $v \in \mathcal{R}_{\Box}(w) \cap \|\neg\varphi\|$. Hence, by **C3** we know that $\mathcal{R}_{[i]}(v) \cap \|\neg\varphi\| \neq \emptyset$. Thus, by Lemma 3.1-(2) we know for all $u \in \mathcal{R}_{[i]}(v), u \models \langle i\rangle\neg\varphi$ and so $\mathfrak{M}, v \models [i]\langle i\rangle\neg\varphi$. Consequently, $\mathfrak{M}, w \models \Diamond[i]\neg[i]\varphi$.

A16 $\otimes_i\varphi \to (\Diamond\varphi \land \Diamond\neg\varphi)$. Combine the reasoning for A11 and A14 with **O6**.

A17 $\otimes_i\varphi \to (\Diamond[i]\varphi \land \Diamond\varphi \land \Diamond\neg\varphi)$. Combine the reasoning for A13 and A16 with **O7**.

A18 $\otimes_i\varphi \to (\Diamond[i]\varphi \land \Diamond[i]\neg\varphi)$. Assume $\mathfrak{M}, w \models \otimes_i\varphi$. The first conjunct follows from similar reasoning as for A13 using **O8**. For the second conjunct, by the semantic definition of $\otimes_i$ we know $\|\varphi\| \in \mathcal{N}_{\otimes_i}(w)$ and so, by property **O8**, we know there is a $v \in \mathcal{R}_{\Box}(w)$ such that $\mathcal{R}_{[i]}(v) \subseteq \overline{\|\varphi\|}$, which by Lemma 3.1-(5) gives us $\mathcal{R}_{[i]}(v) \subseteq \|\neg\varphi\|$. By Lemma 3.1-(3), $\mathfrak{M}, v \models [i]\varphi$ and so $\mathfrak{M}, w \models \Diamond[i]\neg\varphi$. By the semantic definition of $\land$ we have $\mathfrak{M}, w \models \Diamond[i]\varphi \land \Diamond[i]\neg\varphi$.

A19 $\otimes_i\varphi \to \otimes_i\Diamond\varphi$. Assume $\mathfrak{M}, w \models \otimes_i\varphi$. By the semantic definition of $\otimes_i$ we know $\|\varphi\| \in \mathcal{N}_{\otimes_i}(w)$ and so, by property **O9**, we know $\{v \in W \mid \mathcal{R}_{\Box}(v) \cap \|\varphi\| \neq \emptyset\} \in \mathcal{N}_{\otimes_i}(w)$. By Lemma 3.1-(2) and the definition of a truth set we know $\|\Diamond\varphi\| = \{v \in W \mid \mathcal{R}_{\Box}(v) \cap \|\varphi\| \neq \emptyset\} \in \mathcal{N}_{\otimes_i}(w)$. By the semantic definition of $\otimes_i$, we have $\mathfrak{M}, w \models \otimes_i\Diamond\varphi$.

A20 $\otimes_i\varphi \to \otimes_i\Diamond[i]\varphi$. Assume $\mathfrak{M}, w \models \otimes_i\varphi$. By the semantic definition of $\otimes_i$ we know $\|\varphi\| \in \mathcal{N}_{\otimes_i}(w)$ and so, by property **O10**, we know $\Sigma = \{v \in W \mid$ there is a $u \in \mathcal{R}_{\Box}(v)$ such that $\mathcal{R}_{[i]}(u) \subseteq \|\varphi\|\} \in \mathcal{N}_{\otimes_i}(w)$. By Lemma 3.1-(2)-(3), and the definition of a truth set, we have $\Sigma = \|\Diamond[i]\varphi\| \in \mathcal{N}_{\otimes_i}(w)$. By the semantic definition of $\otimes_i$ we have $\mathfrak{M}, w \models \otimes_i\Diamond[i]\varphi$.

R2 Assume $\mathfrak{M} \models \varphi \equiv \psi$. This means $\|\varphi\| = \|\psi\|$. Hence, $\|\varphi\| \in \mathcal{N}_{\otimes_i}(w)$ if and only if $\|\psi\| \in \mathcal{N}_{\otimes_i}(w)$. Therefore, $\mathfrak{M} \models \otimes_i\varphi \equiv \otimes_i\psi$.

The above holds for each $i \in$ Agents. This finishes the proof of (†). The main claim follows from (†) in the usual way (cf. Theorem 2.1). QED

In order to prove strong completeness for the class of $\mathsf{OS}_n\mathsf{X}$ logics, we adapt the method of canonical models for non-normal modal logics (Chellas, 1980). The strategy is as follows: Let $\mathsf{OS}_n\mathsf{X}$ be a logic from Definition 3.3. First, we define the notion of a $\mathsf{OS}_n\mathsf{X}$-maximally consistent set of $\mathcal{L}_n^d$ formulae (Definition 3.7). These sets are used as worlds in the construction of models that are canonical for the logic $\mathsf{OS}_n\mathsf{X}$ (Definition 3.8). Subsequently, we prove a truth lemma (Lemma 3.7), ensuring that every $\mathsf{OS}_n\mathsf{X}$-consistent set of formulae can be satisfied on the corresponding canonical model. The main aim is to demonstrate that the obtained canonical model is an $\mathsf{OS}_n\mathsf{X}$-model (Lemma 3.8). Finally, the model is used to prove completeness via contraposition (Theorem 3.2).

First, we define $\mathsf{OS}_n\mathsf{X}$-consistent sets and $\mathsf{OS}_n\mathsf{X}$-maximally consistent sets.

Definition 3.7 ($\mathsf{OS}_n\mathsf{X}$-CS and $\mathsf{OS}_n\mathsf{X}$-MCS)**.** *Let $\mathsf{OS}_n\mathsf{X}$ be a logic from Definition 3.3. A set $\Delta \subset \mathcal{L}_n^d$ is an* $\mathsf{OS}_n\mathsf{X}$-consistent set *(for short, $\mathsf{OS}_n\mathsf{X}$-CS) iff $\Delta \nvdash_{\mathsf{OS}_n\mathsf{X}} \bot$. A set $\Delta \subset \mathcal{L}_n^d$ is an* $\mathsf{OS}_n\mathsf{X}$-maximally consistent set *(for short, $\mathsf{OS}_n\mathsf{X}$-MCS) iff Δ is an $\mathsf{OS}_n\mathsf{X}$-CS and for any set $\Delta' \subseteq \mathcal{L}_n^d$ such that $\Delta \subset \Delta'$ it is the case that $\Delta' \vdash_{\mathsf{OS}_n\mathsf{X}} \bot$.*

$\mathsf{OS}_n\mathsf{X}$-MCSs have some useful properties. The following Lemma is an adaptation of a general Lemma in (Blackburn et al., 2004). We use these properties implicitly throughout this section.

Lemma 3.2 (Properties of MCSs)**.** *Let $\mathsf{OS}_n\mathsf{X}$ be a logic from Definition 3.3. Let $\Delta \subseteq \mathcal{L}_n^d$ be an $\mathsf{OS}_n\mathsf{X}$-MCS and $\varphi \in \mathcal{L}_n^d$. The following holds:*

- *$\Delta \vdash_{\mathsf{OS}_n\mathsf{X}} \varphi$ iff $\varphi \in \Delta$;*
- *$\varphi \in \Delta$ iff $\neg\varphi \notin \Delta$;*
- *$\varphi \wedge \psi \in \Delta$ iff $\varphi \in \Delta$ and $\psi \in \Delta$.*

Proof. The proof is identical to that of Lemma 2.2. QED

Adapting Lindenbaum's Lemma, every $\mathsf{OS}_n\mathsf{X}$-CS can be extended to an $\mathsf{OS}_n\mathsf{X}$-MCS.

Lemma 3.3 (Lindenbaum's Lemma for $\mathsf{OS}_n\mathsf{X}$)**.** *Let $\mathsf{OS}_n\mathsf{X}$ be a logic from Definition 3.3. Let $\Delta \subseteq \mathcal{L}_n^d$ be an $\mathsf{OS}_n\mathsf{X}$-CS: there is an $\mathsf{OS}_n\mathsf{X}$-MCS $\Delta' \subseteq \mathcal{L}_n^d$ such that $\Delta \subseteq \Delta'$.*

Proof. See (Blackburn et al., 2004, Lem. 4.17) for a general proof. QED

Definition 3.8 (Canonical model for $\mathsf{OS}_n\mathsf{X}$)**.** *Let $\mathsf{OS}_n\mathsf{X}$ be a logic from Definition 3.3. Let $[\alpha] \in \mathsf{Boxes} = \{\Box\} \cup \{[i] \mid i \in \mathsf{Agents}\}$ and let $\langle\alpha\rangle$ be the operator dual to $[\alpha]$. We define the* canonical model *for the logic $\mathsf{OS}_n\mathsf{X}$ to be the tuple $\mathfrak{M}^c := \langle W^c, \mathcal{R}^c_\Box, \{\mathcal{R}^c_{[i]} \mid i \in \mathsf{Agents}\}, \{\mathcal{N}^c_{\otimes_i} \mid i \in \mathsf{Agents}\}, V^c\rangle$ such that:*

- *$W^c := \{\Gamma \subset \mathcal{L}_n^d \mid \Gamma$ is an $\mathsf{OS}_n\mathsf{X}$-MCS$\}$;*
- *for each $[\alpha] \in \mathsf{Boxes}$ and each $\Delta \in W^c$, $\mathcal{R}^c_{[\alpha]}(\Delta) := \{\Gamma \in W^c \mid$ for all $[\alpha]\varphi \in \Delta$, $\varphi \in \Gamma\}$;*

Let $\{\!|\varphi|\!\}_{\mathfrak{M}^c} = \{\Gamma \in W^c \mid \varphi \in \Gamma\}$ be the proof set[12] *of φ (we omit the subscript $\mathfrak{M}^c$):*

- *for each $\Delta \in W^c$, $\mathcal{N}^c_{\otimes_i}(\Delta) := \{Z \subseteq W^c \mid \otimes_i \varphi \in \Delta, \{\!|\varphi|\!\} = Z\}$;*
- *V^c is a valuation function such that for all $p \in \mathsf{Atoms}$, $V^c(p) := \{\Delta \in W^c \mid p \in \Delta\}$.*

[12] It can be straightforwardly shown that proof sets satisfy $\overline{\{\!|\varphi|\!\}} = \{\!|\neg\varphi|\!\}$ and $\{\!|\varphi \wedge \psi|\!\} = \{\!|\varphi|\!\} \cap \{\!|\psi|\!\}$.

The semantic evaluation of $\mathcal{L}_n^d$ formulae on $\mathfrak{M}^c$ is defined as in Definition 3.5.

We show that the defined canonical model possesses certain properties helpful in demonstrating that the canonical model belongs to the class of $\mathsf{OS}_n\mathsf{X}$-models (Lemma 3.8).

Lemma 3.4. *The canonical model $\mathfrak{M}^c$ of Definition 3.8 is well-defined.*

Proof. Through the construction of a simple $\mathsf{OS}_n\mathsf{X}$-model, it is straightforward to show that the logic $\mathsf{OS}_n\mathsf{X}$ is consistent, and so there exists at least one $\mathsf{OS}_n\mathsf{X}$-MCS Γ, such that $\Gamma \in W^c$, i.e., $W^c \neq \emptyset$. By a quick comparison with Definition 3.6 it can be seen that $\mathcal{R}^c_{[\alpha]} \subseteq W^c \times W^c$ and $V^c(p) \subseteq W^c$ for each $p \in \mathsf{Atoms}$. It remains to show that the definition of $\mathcal{N}^c_{\otimes_i}$, defined relative to proofsets only, is not ambiguous. That is, we show that for each $\psi, \varphi \in \mathcal{L}$ and $\Delta \in W^c$,

$$\text{if } \{\!|\varphi|\!\} = \{\!|\psi|\!\}, \text{ then } \{\!|\varphi|\!\} \in \mathcal{N}^c_{\otimes_i}(\Delta) \text{ iff } \{\!|\psi|\!\} \in \mathcal{N}^c_{\otimes_i}(\Delta).$$

Assume that $\{\!|\varphi|\!\} = \{\!|\psi|\!\}$. Hence, $\vdash_{\mathsf{OS}_n\mathsf{X}} \varphi \equiv \psi$ and so by $\mathsf{R2}$ $\vdash_{\mathsf{OS}_n\mathsf{X}} \otimes_i\varphi \equiv \otimes_i\psi$. By the properties of $\mathsf{OS}_n\mathsf{X}$-MCSs, $\otimes_i\varphi \in \Delta$ iff $\otimes_i\psi \in \Delta$ for each $\Delta \in W^c$. Therefore, by the definition of $\mathcal{N}^c_{\otimes_i}$ (Definition 3.8), $\{\!|\varphi|\!\} \in \mathcal{N}^c_{\otimes_i}(\Delta)$ iff $\{\!|\psi|\!\} \in \mathcal{N}^c_{\otimes_i}(\Delta)$. QED

From the above, it follows immediately that for each $\Delta \in W^c$,

$$\otimes_i\varphi \in \Delta \text{ iff } \{\!|\varphi|\!\} \in \mathcal{N}^c_{\otimes_i}(\Delta).$$

Consequently, we know that the $\mathsf{OS}_n\mathsf{X}$ rule $\mathsf{R2}$ is valid on the canonical model for $\mathsf{OS}_n\mathsf{X}$. Next, the existence lemma holds for $\mathsf{OS}_n\mathsf{X}$.

Lemma 3.5 (Existence Lemma $\Box$ and $[i]$)**.** *Let $\mathfrak{M}^c$ be the canonical model. For any world $\Delta \in W^c$ of $\mathfrak{M}^c$ and each $i \in \mathsf{Agents}$ the following holds:*

- *If $\Diamond\varphi \in \Delta$, then there is a $\Gamma \in W^c$ such that $\varphi \in \Gamma$ and $\Gamma \in \mathcal{R}^c_{\Box}(\Delta)$;*
- *If $\langle i\rangle\varphi \in \Delta$, then there is a $\Gamma \in W^c$ such that $\varphi \in \Gamma$ and $\Gamma \in \mathcal{R}^c_{[i]}(\Delta)$.*

Proof. See (Blackburn et al., 2004, Lem. 4.20) for a general proof. QED

Corollary 3.1. *Let $\mathfrak{M}^c$ be the canonical model. For any world $\Delta \in W^c$ of $\mathfrak{M}^c$ and each $i \in \mathsf{Agents}$ the following holds:*

- *If for all $\Gamma \in \mathcal{R}^c_{\Box}(\Delta), \varphi \in \Gamma$, then $\Box\varphi \in \Delta$;*
- *If for all $\Gamma \in \mathcal{R}^c_{[i]}(\Delta), \varphi \in \Gamma$, then $[i]\varphi \in \Delta$.*

Proof. Suppose not, then for all $\Gamma \in \mathcal{R}^c_{\Box}(\Delta)$, $\varphi \in \Gamma$, but $\Box\varphi \notin \Delta$. Hence, $\neg\Box\varphi \in \Delta$ and consequently $\Diamond\neg\varphi \in \Delta$. By Lemma 3.5, there is a Γ with $\neg\varphi \in \Gamma$. Contradiction. QED

The above existence lemma holds for the normal modal operators of $\mathcal{L}_n^d$. For the non-normal modalities $\otimes_i$, we observe the following:

Lemma 3.6 (Existence Lemma $\otimes_i$, (Chellas, 1980))**.** *Let $\mathfrak{M}^c$ be the canonical model. For any world $\Delta \in W^c$ of $\mathfrak{M}^c$ and each $i \in \mathsf{Agents}$ the following holds:*

- $\ominus_i\varphi \in \Delta$ *iff* $\overline{\{|\varphi|\}} \notin \mathcal{N}^c_{\otimes_i}(\Delta)$.

Proof. See (Chellas, 1980, Thm 9.4) for a general proof. QED

The following truth lemma shows that the defined model is canonical for the logic $\mathsf{OS}_n\mathsf{X}$, i.e., each $\mathsf{OS}_n\mathsf{X}$-MCS is satisfiable on this model.

Lemma 3.7 (Truth Lemma)**.** *Let $\mathfrak{M}^c$ be the canonical model. For any $\varphi \in \mathcal{L}_n^d$ and $\Delta \in W^c$ of $\mathfrak{M}^c$: $\mathfrak{M}^c, \Delta \models \varphi$ iff $\varphi \in \Delta$.*

Proof. The proof is by induction on the complexity of φ. It is identical to the truth lemma (Lemma 2.5) proven in Chapter 2. The only exception is the case for $\otimes_i$.

($\varphi = \otimes_i\psi$) We prove the two directions simultaneously. $\mathfrak{M}^c, \Delta \models \otimes_i\psi$ iff $\|\psi\| \in \mathcal{N}^c_{\otimes_i}(\Delta)$ iff $\{\Gamma \in W^c \mid \mathfrak{M}^c, \Gamma \models \psi\} \in \mathcal{N}^c_{\otimes_i}(\Delta)$ iff, by IH, $\{\Gamma \in W^c \mid \psi \in \Gamma\} \in \mathcal{N}^c_{\otimes_i}(\Delta)$ iff $\{|\psi|\} \in \mathcal{N}^c_{\otimes_i}(\Delta)$ iff $\otimes_i\psi \in \Delta$.

The above holds for each $i \in \mathsf{Agents}$. QED

A direct consequence of Lemma 3.7 is that the notion of a truth set coincides with its syntactic counterpart, the notion of a proof set, i.e., $\|\varphi\|_{\mathfrak{M}^c} = \{|\varphi|\}_{\mathfrak{M}^c}$. We use this fact in demonstrating that the canonical model $\mathfrak{M}^c$ is an $\mathsf{OS}_n\mathsf{X}$-model. Due to the modularity of our approach, it suffices to present a single proof for the entire class of $\mathsf{OS}_n\mathsf{X}$ logics.

Lemma 3.8 (Canonical $\mathsf{OS}_n\mathsf{X}$-model)**.** *Let $\mathfrak{M}^c$ be the canonical model: $\mathfrak{M}^c$ belongs to the class of $\mathsf{OS}_n\mathsf{X}$-models.*

Proof. We observe that the cases for the non-deontic frame properties **C1**–**C4** are as for Lemma 2.6 of Section 2.2. We make use of these properties in the cases below. First, $\mathfrak{M}^c$ is well-defined by Lemma 3.4. We only need to show the properties **O1**–**O10**. As we did for proving soundness, we assume that when we demonstrate, for instance, property **O3** that $\mathsf{A12} \in \mathsf{X}$ of $\mathsf{OS}_n\mathsf{X}$ according to Table 3.2. Take an arbitrary $\Delta \in W^c$ of $\mathfrak{M}^c$:

O1 Suppose $Z \in \mathcal{N}^c_{\otimes_i}(\Delta)$ and $\Gamma \in \mathcal{R}^c_{\Box}(\Delta)$. By construction of $\mathfrak{M}^c$ there is an $\otimes_i\varphi \in \Delta$ such that $\{|\varphi|\} = Z$. By the fact that Δ is an $\mathsf{OS}_n\mathsf{X}$-MCS, the axiom $\mathsf{A10}$ $\otimes_i\varphi \to \Box\otimes_i\varphi \in \Delta$ and therefore $\Box\otimes_i\varphi \in \Delta$. By assumption $\Gamma \in \mathcal{R}^c_{\Box}(\Delta)$ and by definition of $\mathcal{R}^c_{\Box}$, $\otimes_i\varphi \in \Gamma$ and so $Z \in \mathcal{N}^c_{\otimes_i}(\Gamma)$ too.

O2 Suppose $Z \in \mathcal{N}^c_{\otimes_i}(\Delta)$. By the construction of $\mathfrak{M}^c$, there is an $\otimes_i\varphi \in \Delta$ such that $\{|\varphi|\} = Z$. By the fact that Δ is an $\mathsf{OS}_n\mathsf{X}$-MCS, the axiom A11 $\otimes_i\varphi \to \neg \otimes_i \bot \in \Delta$ and therefore $\ominus_i\top \in \Delta$. By Lemma 3.6, $\overline{\{|\top|\}} \notin \mathcal{N}^c_{\otimes_i}(\Delta)$ and since $\{|\top|\} = W$, $\overline{\{|\top|\}} = \emptyset$. Consequently, $\emptyset \notin \mathcal{N}^c_{\otimes_i}(\Delta)$.

O3 Suppose $Z \in \mathcal{N}^c_{\otimes_i}(\Delta)$. By the construction of $\mathfrak{M}^c$, there is an $\otimes_i\varphi \in \Delta$ such that $\{|\varphi|\} = Z$. By the fact that Δ is an $\mathsf{OS}_n\mathsf{X}$-MCS, the axiom A12 $\otimes_i\varphi \to \Diamond\varphi \in \Delta$ and therefore $\Diamond\varphi \in \Delta$. By Lemma 3.5, there is a $\Gamma \in W^c$ such that $\Gamma \in \mathcal{R}^c_{\Box}(\Delta)$ and $\varphi \in \Gamma$. Hence, $\Gamma \in \mathcal{R}^c_{\Box}(\Delta) \cap Z \neq \emptyset$.

O4 Suppose $Z \in \mathcal{N}^c_{\otimes_i}(\Delta)$. By the construction of $\mathfrak{M}^c$, there is an $\otimes_i\varphi \in \Delta$ such that $\{|\varphi|\} = Z$. By the fact that Δ is an $\mathsf{OS}_n\mathsf{X}$-MCS, the axiom A13 $\otimes_i\varphi \to \Diamond[i]\varphi \in \Delta$ and therefore $\Diamond[i]\varphi \in \Delta$. By Lemma 3.5, there is a $\Gamma \in W^c$ such that $\Gamma \in \mathcal{R}^c_{\Box}(\Delta)$ and $[i]\varphi \in \Gamma$. Take an arbitrary $\Sigma \in \mathcal{R}^c_{[i]}(\Gamma)$, by definition of $\mathcal{R}^c_{[i]}$, $\varphi \in \Sigma$ and so $\mathcal{R}^c_{[i]}(\Gamma) \subseteq Z$.

O5 Suppose $Z \in \mathcal{N}^c_{\otimes_i}(\Delta)$. By the construction of $\mathfrak{M}^c$, there is an $\otimes_i\varphi \in \Delta$ such that $\{|\varphi|\} = Z$. By the fact that Δ is an $\mathsf{OS}_n\mathsf{X}$-MCS, the axiom A14 $\otimes_i\varphi \to \Diamond\neg\varphi \in \Delta$ and therefore $\Diamond\neg\varphi \in \Delta$. By Lemma 3.5, there is a $\Gamma \in W^c$ such that $\Gamma \in \mathcal{R}^c_{\Box}(\Delta)$ and $\neg\varphi \in \Gamma$. By Lemma 3.1-(5), $\{|\neg\varphi|\} = \overline{\{|\varphi|\}}$, and so $\Gamma \in \mathcal{R}^c_{\Box}(\Delta) \cap \overline{Z} \neq \emptyset$.[13]

O6 Suppose $Z \in \mathcal{N}^c_{\otimes_i}(\Delta)$. By the construction of $\mathfrak{M}^c$, there is an $\otimes_i\varphi \in \Delta$ such that $\{|\varphi|\} = Z$. By the fact that Δ is an $\mathsf{OS}_n\mathsf{X}$-MCS, the axiom A16 $\otimes_i\varphi \to (\Diamond\varphi \wedge \Diamond\neg\varphi) \in \Delta$ and therefore $\Diamond\varphi, \Diamond\neg\varphi \in \Delta$. Proceed as for **O3** and **O5**.

O7 Combine the reasoning for **O4** and **O6** with axiom A17.

O8 Suppose $Z \in \mathcal{N}^c_{\otimes_i}(\Delta)$. By the construction of $\mathfrak{M}^c$, there is an $\otimes_i\varphi \in \Delta$ such that $\{|\varphi|\} = Z$. By the fact that Δ is an $\mathsf{OS}_n\mathsf{X}$-MCS, the axiom A18 $\otimes_i\varphi \to (\Diamond[i]\varphi \wedge \Diamond[i]\neg\varphi) \in \Delta$. Therefore, $\Diamond[i]\varphi \wedge \Diamond[i]\neg\varphi \in \Delta$ and so $\Diamond[i]\varphi, \Diamond[i]\neg\varphi \in \Delta$. For the first conjunct, proceed as for **O4**. For the second conjunct, by Lemma 3.5 there is a $\Gamma \in W^c$ such that $\Gamma \in \mathcal{R}^c_{\Box}(\Delta)$ and $[i]\neg\varphi \in \Gamma$. Take an arbitrary $\Sigma \in \mathcal{R}^c_{[i]}(\Gamma)$, by definition of $\mathcal{R}^c_{[i]}$, $\neg\varphi \in \Sigma$ and so $\mathcal{R}^c_{[i]}(\Gamma) \subseteq \overline{Z} = \{|\neg\varphi|\}$.

O9 Suppose $Z \in \mathcal{N}^c_{\otimes_i}(\Delta)$. By the construction of $\mathfrak{M}^c$, there is an $\otimes_i\varphi \in \Delta$ such that $\{|\varphi|\} = Z$. By the fact that Δ is an $\mathsf{OS}_n\mathsf{X}$-MCS, the axiom A19 $\otimes_i\varphi \to \otimes_i\Diamond\varphi \in \Delta$ and therefore $\otimes_i\Diamond\varphi \in \Delta$. By definition of $\mathcal{N}^c_{\otimes_i}$ we know $\{|\Diamond\varphi|\} \in \mathcal{N}^c_{\otimes_i}(\Delta)$. By $\{|\Diamond\varphi|\} = \{\Gamma \in W^c \mid \Diamond\varphi \in \Gamma\}$ and Lemma 3.5 and the definition of $\mathcal{R}^c_{\Box}$, $\Diamond\varphi \in \Gamma$ iff there is a $\Sigma \in W^c$ such that $\varphi \in \Sigma$ and $\Sigma \in \mathcal{R}^c_{\Box}(\Gamma)$. Consequently, $\{|\Diamond\varphi|\} = \{\Gamma \in W^c \mid \varphi \in \Sigma \text{ and } \Gamma \in \mathcal{R}^c_{\Box}(\Sigma)\} = \{\Gamma \in W^c \mid \Sigma \in \{|\varphi|\} \text{ and } \Gamma \in \mathcal{R}^c_{\Box}(\Sigma)\} = \{\Gamma \in W^c \mid \Sigma \in \{|\varphi|\} \text{ and } \Sigma \in \mathcal{R}^c_{\Box}(\Gamma)\} = \{\Gamma \in W^c \mid \Sigma \in \{|\varphi|\} \cap \mathcal{R}^c_{\Box}(\Gamma)\} = \{\Gamma \in W^c \mid Z \cap \mathcal{R}^c_{\Box}(\Gamma) \neq \emptyset\} \in \mathcal{N}^c_{\otimes_i}(\Delta)$.

[13] Similar reasoning applies for the case where the axiom A15 $\otimes_i\varphi \to \Diamond[i]\neg[i]\varphi \in \mathsf{X}$ of $\mathsf{OS}_n\mathsf{X}$. One can make use of the fact that $\vdash_{\mathsf{OS}_n\mathsf{X}} \mathsf{OiV} \equiv \mathsf{OiRef}$ (Remark 3.3).

O10 Suppose $Z \in \mathcal{N}^c_{\otimes_i}(\Delta)$. By the construction of $\mathfrak{M}^c$, there is an $\otimes_i\varphi \in \Delta$ such that $\{|\varphi|\} = Z$. By the fact that Δ is an $\mathsf{OS}_n\mathsf{X}$-MCS, the axiom $\mathsf{A20}$ $\otimes_i\varphi \to \otimes_i\Diamond[i]\varphi \in \Delta$ and therefore $\otimes_i\Diamond[i]\varphi \in \Delta$. By construction of $\mathcal{N}^c_{\otimes_i}$ we know $\{|\Diamond[i]\varphi|\} \in \mathcal{N}^c_{\otimes_i}(\Delta)$. $\{|\Diamond[i]\varphi|\} = \{\Gamma \in W^c \mid \Diamond[i]\varphi \in \Gamma\}$. By Lemma 3.5 and construction of $\mathcal{R}^c_{\Box}$, $\Diamond[i]\varphi \in \Gamma$ iff there is a $\Sigma \in W^c$ such that $[i]\varphi \in \Sigma$ and $\Sigma \in \mathcal{R}^c_{\Box}(\Gamma)$. Consequently, $\{|\Diamond[i]\varphi|\} = \{\Gamma \in W^c \mid [i]\varphi \in \Sigma$ and $\Sigma \in \mathcal{R}^c_{\Box}(\Gamma)\} = \{\Gamma \in W^c \mid \mathcal{R}^c_{[i]}(\Sigma) \subseteq \{|\varphi|\}$ and $\Sigma \in \mathcal{R}^c_{\Box}(\Gamma)\} = \{\Gamma \in W^c \mid \mathcal{R}^c_{[i]}(\Sigma) \subseteq Z$ and $\Sigma \in \mathcal{R}^c_{\Box}(\Gamma)\} \in \mathcal{N}^c_{\otimes_i}(\Delta)$. QED

Since all the above results were shown for an arbitrary logic $\mathsf{OS}_n\mathsf{X}$ from Definition 3.3, we can now demonstrate strong completeness for all $\mathsf{OS}_n\mathsf{X}$ logics.

Theorem 3.2 (Strong Completeness of $\mathsf{OS}_n\mathsf{X}$)**.** *Let $\mathsf{OS}_n\mathsf{X}$ be a logic from Definition 3.3. For any formula $\varphi \in \mathcal{L}^d_n$, and any $\Gamma \subseteq \mathcal{L}^d_n$: if $\Gamma \models_{\mathsf{OS}_n\mathsf{X}} \varphi$, then $\Gamma \vdash_{\mathsf{OS}_n\mathsf{X}} \varphi$.*

Proof. The proof is by contraposition. Suppose φ is not $\mathsf{OS}_n\mathsf{X}$-derivable from Γ. This means that $\Gamma \cup \{\neg\varphi\}$ is an $\mathsf{OS}_n\mathsf{X}$-CS. Namely, if $\Gamma \cup \{\neg\varphi\}$ would be $\mathsf{OS}_n\mathsf{X}$-inconsistent, then $\Gamma, \neg\varphi \vdash_{\mathsf{OS}_n\mathsf{X}} \bot$ and so $\Gamma \vdash_{\mathsf{OS}_n\mathsf{X}} \varphi$. By Lemma 3.3 there is a $\Gamma' \subseteq \mathcal{L}^d_n$ such that Γ' is an $\mathsf{OS}_n\mathsf{X}$-MCS and $\Gamma \cup \{\neg\varphi\} \subseteq \Gamma'$. By construction of the canonical model, $\Gamma' \in W^c$ and by Lemma 3.7 we know that $\mathfrak{M}^c, \Gamma' \models \Gamma$ and $\mathfrak{M}^c, \Gamma' \models \neg\varphi$. By Lemma 3.8, $\mathfrak{M}^c$ is an $\mathsf{OS}_n\mathsf{X}$-model and so $\Gamma \not\models_{\mathsf{OS}_n\mathsf{X}} \varphi$. QED

3.4 A Formal Taxonomy of Ought Implies Can

In this section, we put our $\mathsf{OS}_n\mathsf{X}$ logics to work and address Objective 2. First, we organize the logics in terms of their strength: observing which are equivalent, distinct, or subsumed by another. Second, we discuss the logical (in)dependencies between the various OiC principles by comparing the minimal systems in which each principle is validated. In Figure 3.1, we provide a lattice ordering the ten $\mathsf{OS}_n\mathsf{X}$ logics extended with a singular OiC axiom (reflexive and transitive edges are left implicit). An ordering of the entire class of $\mathsf{OS}_n\mathsf{X}$ logics can be obtained in a similar way. Concerning Objective 2, it suffices to consider OS_n logics extended with individual OiC axiom schemes.

We consider a logic $\mathsf{OS}_n\mathsf{X}$ stronger than another logic $\mathsf{OS}_n\mathsf{Y}$ whenever the former generates at least the same set of valid formulae as the latter. In Figure 3.1, the directed edges denote subsumption relations, e.g., OS_n (without any OiC axiom) is the smallest logic subsumed by all others, whereas $\mathsf{OS}_n\{\mathsf{Ai}\}$ for $i \in \{18, 19, 20\}$ are the logics not subsumed by any other logic in the lattice (see Remark 3.4 below). To determine the existence of a directed edge from one logic $\mathsf{OS}_n\mathsf{X}$ to another $\mathsf{OS}_n\mathsf{Y}$ in the lattice, it suffices to show that every valid formula of the former is a valid formula of the latter. As an example, we consider the edge from $\mathsf{OS}_n\{\mathsf{A12}\}$ to $\mathsf{OS}_n\{\mathsf{A11}\}$.

Example 3.1. *Observe that the axiomatizations of $\mathsf{OS}_n\mathsf{X}$ and $\mathsf{OS}_n\mathsf{Y}$ with $\mathsf{X} = \{\mathsf{A12}\}$ and $\mathsf{Y} = \{\mathsf{A11}\}$ only differ on the OiC axioms $\mathsf{A12}$ and $\mathsf{A11}$. We demonstrate that the* *latter axiom is also a valid formula in the logic of the former. We know* $\models_{\mathsf{OS}_n\mathsf{X}} \Box\top$ *by the normality of* $\Box$*, and by duality, we know that* $\models_{\mathsf{OS}_n\mathsf{X}} \neg\Diamond\bot$*. By the (straightforward) validity of modus tollens and* A12*, we know* $\models_{\mathsf{OS}_n\mathsf{X}} \neg\otimes_i\bot$*. Hence,* $\models_{\mathsf{OS}_n\mathsf{X}} \otimes_i\varphi \rightarrow \neg\otimes_i\bot$ *for any* $\otimes_i\varphi \in \mathcal{L}^d_n$*.*

The non-existence of a directed edge in the opposite direction is implied by the fact that $\models_{\mathsf{OS}_n\mathsf{X}} \otimes_i\varphi \rightarrow \Diamond\varphi$ *and* $\not\models_{\mathsf{OS}_n\mathsf{Y}} \otimes_i\varphi \rightarrow \Diamond\varphi$*. For the latter claim it suffices to construct an* $\mathsf{OS}_n\mathsf{Y}$ *counter-model: Let* $\mathfrak{M} = \langle W, \mathcal{R}_\Box, \{\mathcal{R}_{[i]} \mid i \in \mathsf{Agents}\}, \{\mathcal{N}_{\otimes_i} \mid i \in \mathsf{Agents}\}, V\rangle$ *where* $W = \{w, v\}, V(p) = \{v\}$ *for each* $p \in \mathsf{Atoms}$*,* $\mathcal{R}_\Box = \{(w, w), (v, v)\} = \mathcal{R}_{[i]}$ *for each* $i \in \mathsf{Agents}$*, and* $\mathcal{N}_{\otimes_i}(w) = \{v\}$ *and* $\mathcal{N}_{\otimes_i}(v) = \{w\}$ *for each* $i \in \mathsf{Agents}$*. It can be easily checked that* $\mathfrak{M}$ *is an* $\mathsf{OS}_n\mathsf{Y}$ *model with* $k \geq 1$ *and* $\mathfrak{M}, w \models \otimes_i p \wedge \neg\Diamond p$*.*

To determine that two logics $\mathsf{OS}_n\mathsf{X}$ and $\mathsf{OS}_n\mathsf{Y}$ are equivalent—i.e., $\mathsf{OS}_n\mathsf{X} = \mathsf{OS}_n\mathsf{Y}$—one shows that every valid formula of the former is a valid formula of the latter, and vice versa. Last, to prove that two logics $\mathsf{OS}_n\mathsf{X}$ and $\mathsf{OS}_n\mathsf{Y}$ are independent—i.e., yielding incomparable logics—it is sufficient to show that there exist formulae φ and ψ such that $\models_{\mathsf{OS}_n\mathsf{X}} \varphi$, $\not\models_{\mathsf{OS}_n\mathsf{Y}} \varphi$, $\models_{\mathsf{OS}_n\mathsf{Y}} \psi$, and $\not\models_{\mathsf{OS}_n\mathsf{X}} \psi$. To illustrate this, we consider the logics containing A19 and A14.

Example 3.2. *For brevity, assume a single-agent setting for which* $\{i\} = \mathsf{Agents}$*. Let* $\mathsf{OS}_n\mathsf{X}$ *and* $\mathsf{OS}_n\mathsf{Y}$ *be such that* $\mathsf{X} = \{\mathsf{A19}\}$ *and* $\mathsf{Y} = \{\mathsf{A14}\}$*. We know that* $\models_{\mathsf{OS}_n\mathsf{X}} \otimes_i p \rightarrow \otimes_i\Diamond p$ *and* $\models_{\mathsf{OS}_n\mathsf{Y}} \otimes_i p \rightarrow \Diamond\neg p$ *for each* $p \in \mathsf{Atoms}$*. It suffices to provide counter-models that show* $\not\models_{\mathsf{OS}_n\mathsf{X}} \otimes_i p \rightarrow \Diamond\neg p$ *and* $\not\models_{\mathsf{OS}_n\mathsf{Y}} \otimes_i p \rightarrow \otimes_i\Diamond p$ *for some* $p \in \mathsf{Atoms}$*. Let* $\mathfrak{M}^j = \langle W, \mathcal{R}^j_\Box, \mathcal{R}^j_{[i]}, \mathcal{N}^j_{\otimes_i}, V^j\rangle$ *with* $j \in \{1, 2\}$*. First, let* $W^1 = \{w, v\}$*,* $\mathcal{R}^1_\Box = \{(w, w), (v, v), (w, v), (v, w)\}$*,* $\mathcal{R}^1_{[i]} = \{(w, w), (v, v)\}$*,* $V^1(p) = \{v\}$ *for each* $p \in \mathsf{Atoms}$*, and* $\mathcal{N}^1_{\otimes_i}(w) = \{v\} = \mathcal{N}^1_{\otimes_i}(v)$*. We have* $\mathfrak{M}^1, w \models \otimes_i p \wedge \neg\otimes_i\Diamond p$*, since* $\|\Diamond p\| = \{w, v\} \notin \mathcal{N}^1_{\otimes_i}(w)$*, and* $\mathfrak{M}^1$ *is an* $\mathsf{OS}_n\mathsf{Y}$*-model. Second, let* $W^2 = \{w\}, \mathcal{R}^2_\Box = \{(w, w)\} = \mathcal{R}^2_{[i]}$*,* $V^2(p) = \{w\}$ *for each* $p \in \mathsf{Atoms}$*, and* $\mathcal{N}^2_{\otimes_i}(w) = \{w\}$*. We have* $\mathfrak{M}^2, w \models \otimes_i p \wedge \Box p$ *and* $\mathfrak{M}^2$ *is an* $\mathsf{OS}_n\mathsf{X}$ *model. The models* $\mathfrak{M}^1$ *and* $\mathfrak{M}^2$ *are graphically depicted below (only relevant formulae are explicitly represented and deontically ideal worlds are shaded).*

$\neg p$
w
p
v
$\mathfrak{M}^1$
$\neg p$
w
$\mathfrak{M}^2$

Excluding the normative readings of OiC—i.e., OiNC and OiNA—we may say that OiCtrl is the strongest OiC principle since it entails all other OiC principles. Furthermore, all OiC principles are compatible with each other. That is, any combination of OiC axioms generates a consistent logic. This can be straightforwardly checked. Last, the taxonomy shows that both OiNC and OiNA are strictly independent of any other OiC principle.

Remark 3.4. *In (van Berkel and Lyon, 2021) we concluded that* $\otimes_i\varphi \rightarrow \otimes_i\Diamond[i]\varphi$ *(*A20*) is a stronger reading of OiC than* $\otimes_i\varphi \rightarrow \otimes_i\Diamond\varphi$ *(*A19*), the former entailing the latter.*

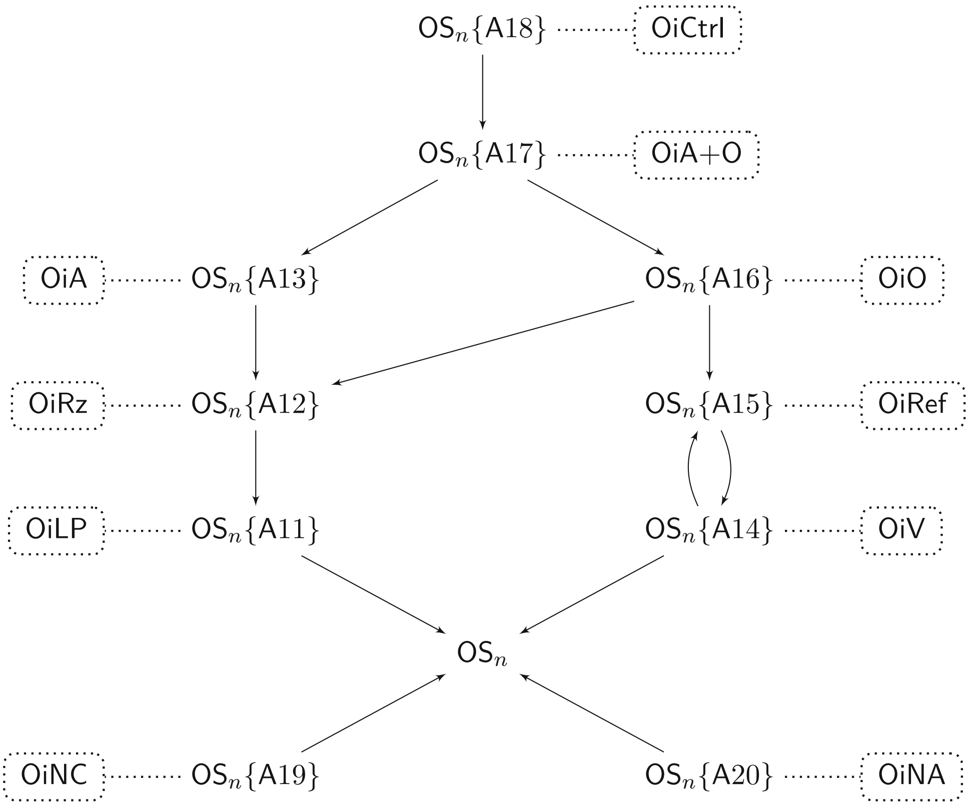

$\mathsf{OS}_n\{\mathsf{A18}\}$
OiCtrl
$\mathsf{OS}_n\{\mathsf{A17}\}$
OiA+O
OiA
$\mathsf{OS}_n\{\mathsf{A13}\}$
$\mathsf{OS}_n\{\mathsf{A16}\}$
OiO
OiRz
$\mathsf{OS}_n\{\mathsf{A12}\}$
$\mathsf{OS}_n\{\mathsf{A15}\}$
OiRef
OiLP
$\mathsf{OS}_n\{\mathsf{A11}\}$
$\mathsf{OS}_n\{\mathsf{A14}\}$
OiV
OS_n
OiNC
$\mathsf{OS}_n\{\mathsf{A19}\}$
$\mathsf{OS}_n\{\mathsf{A20}\}$
OiNA

Figure 3.1: The lattice of $\mathsf{OS}_n\mathsf{X}$ logics for OiC. Directed edges point from stronger logics to weaker logics with respect to their expressivity. Reflexive and transitive edges are left implicit. The two arrows between $\mathsf{OS}_n\{\mathsf{A15}\}$ and $\mathsf{OS}_n\{\mathsf{A14}\}$ denote that the logics are equivalent. The logic OS_n (without any OiC axiom) is subsumed by all others, whereas the logics $\mathsf{OS}_n\{\mathsf{A18}\}$, $\mathsf{OS}_n\{\mathsf{A19}\}$, and $\mathsf{OS}_n\{\mathsf{A20}\}$ are subsumed by no other logic. Dotted nodes make explicit which logic corresponds to which OiC principle.

However, our present analysis—depicted in Figure 3.1—concludes that the two OiC interpretations are independent. We are now in a better position to understand the two. In (van Berkel and Lyon, 2021), $\otimes_i$ was characterized as a normal modal operator. Since $\Diamond[i]\varphi \rightarrow \Diamond\varphi$ is a theorem of an any normal modal deontic extension of basic STIT *logic (e.g.,* DS_n *and* US_n*) we can prove that $\otimes_i\varphi \rightarrow \otimes_i\Diamond[i]\varphi$ implies $\otimes_i\varphi \rightarrow \otimes_i\Diamond\varphi$ (in the logic in question). The formal proof is omitted. In other words, the dependence concluded in our previous work was due to the endorsement of deontic properties other than OiC, i.e., the properties* M, C, *and* N *for $\otimes_i$ (see page 88). The added value of the approach taken in this chapter is that we can compare OiC readings without endorsing additional properties influencing their interdependencies.*

In this section, it suffices only to consider OS_n logics extended with individual OiC axiom schemes. However, there are further analyses possible. To give an example, the logic $\mathsf{OS}_n\mathsf{X}$ for which $\mathsf{X} = \{\mathsf{A12},\mathsf{A14}\}$ is equivalent to the logic $\mathsf{OS}_n\mathsf{Y}$ with $\mathsf{Y} = \{\mathsf{A16}\}$. In other words, certain combinations of OiC principles correspond to other OiC principles. We omit such an analysis here and return to it in Section 3.5.

From a philosophical perspective, Figure 3.1 gives rise to what we refer to as an *endorsement principle*. Namely, the taxonomy explicates which endorsement of which OiC reading logically commits one to endorsing other OiC readings (from the vantage of STIT). For instance, endorsing OiA tells us that we must also endorse the weaker OiLP and OiRz since they are logically implied in any logic containing OiA.

Definition 3.9 (The Endorsement Principle)**.** *For any two logics* $\mathsf{OS}_n\{\mathsf{A}i\}$ *and* $\mathsf{OS}_n\{\mathsf{A}j\}$*, for* $11 \leq i, j \leq 20$*, if the former subsumes the latter according to Figure 3.1, then an endorsement of the OiC principle* $\mathsf{A}i$ *also commits one to endorsing OiC principle* $\mathsf{A}j$*.*

Alternatively, one may take the endorsement principle in Definition 3.9 to reveal which standpoints are untenable. To illustrate, one cannot endorse OiCtrl as a metaethical principle and at the same time refute OiLP as a valid deontic principle (or any other except for OiNC and OiNA). In the next section, we discuss several other metaethical principles and prove how these logically relate to OiC.

3.5 Other Metaethical Principles and Ought implies Can Reasoning

Now that we have a clearer picture of how the different readings of OiC are logically related to one another in the logic of STIT, we can start exploiting the logical taxonomy to make observations about other metaethical principles and OiC. This is Objective 3. We discuss four such principles:

- No Vacuous Commands (NVC);
- Deontic Contingency (DCg);
- Deontic Consistency (DCs);
- No Deontic Dilemmas (NDD).

Along the way, we prove soundness and strong completeness for the class of $\mathsf{OS}_n\mathsf{X}$ logics extended with the above principles.

Furthermore, we address Objective 4 throughout our discussion. Recall that a common objection to adopting non-normal modal logics is that the logics become too weak and certain intuitively desirable inferences are lost (Van Fraassen, 1973; Horty, 1994). We address Objective 4 and the above objection by considering extensions of $\mathsf{OS}_n\mathsf{X}$ with restricted deontic reasoning principles that take into account OiC. In particular, we consider restricted versions of monotonicity (M) and aggregation (C), as well as a variation of *disjunctive response*.

3.5.1 No Vacuous Commands

This principle, to which we refer as the "No Vacuous Commands" principle (NVC), excludes commands that prescribe states of affairs that will be the case irrespective of an agent's behavior, as well as those which prescribe actions that agents will necessarily perform. The principle is, therefore, closely related to deliberate agency. Following von Wright (1963a), a command to either open a window or keep it closed is satisfied irrespective of what the agent does in that situation, and so "[t]he command, therefore, does not, properly speaking, 'demand' anything at all" (p.153). The NVC principle excludes such obligations. The axiom A21 expresses a version of this principle.

A21. $\neg \otimes_i \top$

Following von Wright (1963a), one may adopt a stronger, agentive interpretation of NVC. Namely, "[t]here is no such thing as making or ('actively') letting people do things which they will necessarily do in any case" (p.154). The following axiom characterizes this last quote:

$$\Box\varphi \rightarrow \neg \otimes_i \varphi$$

This axiom expresses that everything that is currently settled true does not fall within the scope of an obligation. In fact, the formula is nothing but the contraposition of axiom A14 characterizing OiV, i.e., $\vdash_{\mathsf{OS}_n\mathsf{X}} \Box\varphi \rightarrow \neg \otimes_i \varphi \equiv \otimes_i\varphi \rightarrow \Diamond\neg\varphi$ for any $\mathsf{OS}_n\mathsf{X}$. It can be straightforwardly checked that $\Box\varphi \rightarrow \neg \otimes_i \varphi$ implies the axiom A21, i.e., $\neg \otimes_i \top$ (due to the normality of $\Box$).

The class of $\mathsf{OS}_n\mathsf{X}$ logics can be extended with NVC, i.e., the axiom A21. The corresponding frame property is,

O11 For all $w \in W, W \notin \mathcal{N}_{\otimes_i}(w)$

The resulting logics preserve soundness and completeness.

Theorem 3.3. *Any* $\mathsf{OS}_n\mathsf{X}$ *logic extended with* A21 *is sound and complete with respect to its corresponding class of* $\mathsf{OS}_n\mathsf{X}$*-frames extended with* ***O11***.[14]

Proof. Due to the modularity of the soundness and completeness proofs in Section 3.3 it suffices to only consider the additional case for A21 and **O11**. For soundness, we need to prove that A21 is valid on all $\mathsf{OS}_n\mathsf{X}$-models extended with **O11**. For completeness, it

[14] We assume the inclusion of A21 and **O11** for each $i \in \mathsf{Agents}$. We leave this implicit.

suffices to show that the canonical model $\mathfrak{M}^c$—constructed from the logic $\mathsf{OS}_n\mathsf{X}$ extended with $\mathsf{A21}$—satisfies **O11**. [15]

Soundness. Take an arbitrary $\mathsf{OS}_n\mathsf{X}$-model $\mathfrak{M}$ extended with **O11** and let $w \in W$. Since $W \notin \mathcal{N}_{\otimes_i}(w)$ and $W = \|\top\|$, by semantic definition of $\otimes_i$ we have $\mathfrak{M}, w \models \neg \otimes_i \top$.

Completeness. Let $\mathfrak{M}^c$ be a canonical model for the logic $\mathsf{OS}_n\mathsf{X}$ extended with $\mathsf{A21}$. Let $\Delta \in W^c$. By duality, $\neg \otimes_i \top$ is equivalent to $\ominus_i \bot$. By Lemma 3.6 and the fact that $\ominus_i \bot \in \Delta$, we have $\overline{\{\!|\bot|\!\}} \notin \mathcal{N}^c_{\otimes_i}(\Delta)$. Since $\{\!|\bot|\!\} = \emptyset$, we have $\overline{\{\!|\bot|\!\}} = W \notin \mathcal{N}^c_{\otimes_i}(\Delta)$. QED

NVC, expressed as $\neg \otimes_i \top$, is a logical consequence of OiV but not the other way around (as a straightforward $\mathsf{OS}_n\{\mathsf{A21}\}$-countermodel would confirm). The logic $\mathsf{OS}_n\{\mathsf{A21}\}$ can therefore be located in the formal taxonomy of Figure 3.1 between OS_n and $\mathsf{OS}_n\{\mathsf{A14}\}$. Consequently, the endorsement principle tells us that all OiC logics implying OiV (e.g., those containing $\mathsf{OiRef}, \mathsf{OiO}$, and $\mathsf{OiA{+}O}$) commit one to endorse the principle of No Vacuous Commands. As discussed in Section 3.1, these OiC readings require that what is obliged must be voliable, either by mere contingency, the agent's refraining from satisfying the obligation, or by the possibility of the agent seeing to it that the obligation is violated.

The non-deliberative readings of OiC (i.e., $\mathsf{OiLP}, \mathsf{OiRz}, \mathsf{OiA}, \mathsf{OiNC}$, and OiNA) do allow for a consistent modeling of tautologous obligations. Nevertheless, due to the absence of the necessity axiom N in $\mathsf{OS}_n\mathsf{X}$ logics, the formula $\otimes_i \top$ is not a theorem of those logics. In fact, these logics satisfy von Wright's weaker interpretation of NVC as proposed in the seminal work "Deontic Logic" (von Wright, 1951): tautologies are not *necessarily* obligatory. We discuss this interpretation further in Section 3.6. It must be noted that normal modal logics for deontic STIT logic (e.g., DS_n of Chapter 2), trivially violate NVC due to the inclusion of necessitation for $\otimes_i$.

Reasoning with NVC. In case $\otimes_i$ satisfies monotonicity M, whenever a set of assumptions $\Gamma \subseteq \mathcal{L}^d_n$ contains a formula $\otimes_i \varphi$, the obligation $\otimes_i \top$ is implied and NVC is violated. To see this point, let $\vdash$ denote the consequence relation of a $\mathsf{OS}_n\mathsf{X}$ logic satisfying M and let $\otimes_i \varphi \in \Gamma$. Since $\vdash \varphi \equiv (\varphi \wedge \top)$, by $\mathsf{R2}$ we have $\Gamma \vdash \otimes_i(\varphi \wedge \top)$, which by monotonicity of $\otimes_i$ gives us $\Gamma \vdash \otimes_i \varphi \wedge \otimes_i \top$. Moreover, in case the logic in question also contains the $\mathsf{A14}$, we have $\vdash_{\mathsf{OS}_n\mathsf{X}} \otimes_i \varphi \rightarrow (\otimes_i \top \wedge \neg \otimes_i \top)$ (for the second conjunct in the consequent see page 103). Consequently, contraposition gives us $\vdash_{\mathsf{OS}_n\mathsf{X}} \top \rightarrow \neg \otimes_i \varphi$ for each $\varphi \in \mathcal{L}^d_n$. Hence, although such a logic is consistent, nothing is obligatory. In order to have both monotonicity and NVC we must therefore adopt a *restricted* version of axiom M. For instance, consider the following axiom:

$\mathsf{A22}$. $(\Box(\varphi \rightarrow \psi) \wedge \neg \Box \psi \wedge \otimes_i \varphi) \rightarrow \otimes_i \psi$

[15] By their generality, the existence lemmata 3.5 and 3.6 and truth lemma 3.7 hold for all the logics considered in this section.

Axiom $\mathsf{A22}$ enables the derivation of an obligation $\otimes_i\psi$ from another obligation $\otimes_i\varphi$, whenever it is settled true that φ implies ψ and ψ is not vacuously true at the present moment. The corresponding frame property would be:

O12 For all $w \in W, Z, X \subseteq W$, if $Z \in \mathcal{N}_{\otimes_i}(w), \mathcal{R}_\Box(w) \not\subseteq X$, and $\mathcal{R}_\Box(w) \subseteq \overline{Z} \cup X$, then $X \in \mathcal{N}_{\otimes_i}(w)$

Axiom $\mathsf{A22}$ is compatible with all OiC readings. Nevertheless, from a conceptual point of view, it is doubtful whether an axiom like $\mathsf{A22}$ must be adopted in combination with OiNC and OiNA. Namely, these readings of OiC deliberately do not refer to the moment of evaluation. Therefore, a principle expressing monotonicity *restricted to the present moment* is unsuitable in this context. In what follows, we refrain from adding such a principle for normative OiC.

Proving soundness and completeness for the resulting logics requires additional machinery that accounts for monotonicity in non-normal modal logics (Chellas, 1980).

Theorem 3.4. *Let* $\mathsf{X} \subseteq \{\mathsf{A}i \mid 11 \leq i \leq 18\}$. *Any* $\mathsf{OS}_n\mathsf{X}$ *logic extended with axiom* $\mathsf{A22}$ *is sound and complete with respect to their corresponding class of* $\mathsf{OS}_n\mathsf{X}$*-frames extended with* ***O12****.*

Proof. Soundness. Due to the modularity of the soundness proof in Section 3.3, it suffices to only consider the additional case for $\mathsf{A22}$. Take an arbitrary $\mathsf{OS}_n\mathsf{X}$-model $\mathfrak{M}$ extended with **O12** and let $w \in W$. Assume $\mathfrak{M}, w \models \Box(\varphi \to \psi) \wedge \neg\Box\psi \wedge \otimes_i\varphi$. By the semantic definitions of $\otimes_i$ and $\Box$, $\|\varphi\| \in \mathcal{N}_{\otimes_i}(w)$ and there is a $v \in \mathcal{R}_\Box(w)$ such that $\mathfrak{M}, v \models \neg\psi$, i.e., $\mathcal{R}_\Box(w) \not\subseteq \|\psi\|$. Furthermore, $\mathfrak{M}, w \models \Box(\varphi \to \psi)$ if and only if for all $u \in \mathcal{R}_\Box(w)$, if $u \in \|\varphi\|$ then $u \in \|\psi\|$, which is equivalent to $u \in \overline{\|\varphi\|} \cup \|\psi\|$. In other words, $\mathcal{R}_\Box(w) \subseteq \overline{\|\varphi\|} \cup \|\psi\|$. By **O12**, $\|\psi\| \in \mathcal{N}_{\otimes_i}(w)$ and so $\mathfrak{M}, w \models \otimes_i\psi$.

Completeness. So far we considered only *smallest* canonical models $\mathfrak{M}^c$, i.e., where $\mathcal{N}^c_{\otimes_i}(\Delta)$ consists of only those proof sets $\{\!|\varphi|\!\}$ for which $\otimes_i\varphi \in \Delta$ (see page 96). In order to characterize restricted monotonicity, we must extend $\mathcal{N}^c_{\otimes_i}$ with a specific collection of *non*-proof sets. Non-proof sets are sets of MCSs that do not characterize a specific formula from $\mathcal{L}^d_n$ (Chellas, 1980, Ch.9). We show that extensions with non-proof sets preserve the canonicity of the resulting model. We use *supplemented* models (Chellas, 1980, Ch.9), which suffice for all logics extended with axioms from $\{\mathsf{A}i \mid 11 \leq i \leq 18\}$.

Let $\mathfrak{M}^c = \langle W^c, \mathcal{R}^c_\Box, \{\mathcal{R}^c_{[i]} \mid i \in \mathsf{Agents}\}, \{\mathcal{N}^c_{\otimes_i} \mid i \in \mathsf{Agents}\}, V^c\rangle$ be the canonical model as defined in Definition 3.8. The *supplement* canonical model for the logic $\mathsf{OS}_n\mathsf{X}$ extended with $\mathsf{A22}$ is the tuple $\mathfrak{M}^{c,+} = \langle W^c, \mathcal{R}^c_\Box, \{\mathcal{R}^c_{[i]} \mid i \in \mathsf{Agents}\}, \{\mathcal{N}^{c,+}_{\otimes_i} \mid i \in \mathsf{Agents}\}, V^c\rangle$. Where for each $\Delta \in W^c$,

$$\mathcal{N}^{c,+}_{\otimes_i}(\Delta) = \mathcal{N}^c_{\otimes_i}(\Delta) \cup \mathcal{N}^+_{\otimes_i}(\Delta)$$

with

$$\mathcal{N}^+_{\otimes_i}(\Delta) = \{Z \mid \otimes_i\varphi \in \Delta,\ \text{there is a } \Gamma \in \mathcal{R}^c_\Box(\Delta) \text{ s.t. } \mathcal{R}^c_{[i]}(\Gamma) \subseteq \overline{Z} \text{ and } \mathcal{R}^c_\Box(\Delta) \subseteq \overline{\{\!|\varphi|\!\}} \cup Z\}$$

We show that the supplemented model $\mathfrak{M}^{c,+}$ is a canonical model for $\mathsf{OS}_n\mathsf{X}$ extended with A22. For this, it suffices to show the following:

$$\otimes_i\psi \in \Delta \text{ iff } \{|\psi|\} \in \mathcal{N}^{c,+}_{\otimes_i}(\Delta)$$

Left-to-right. It suffices to show $\mathcal{N}^{c}_{\otimes_i}(\Delta) \subseteq \mathcal{N}^{c,+}_{\otimes_i}(\Delta)$, which holds trivially.

Right-to-left. Assume some $Z = \{|\psi|\} \in \mathcal{N}^{c,+}_{\otimes_i}(\Delta)$ such that for some $\otimes_i\varphi \in \Delta$, there is a $\Gamma \in \mathcal{R}^c_{\Box}(\Delta)$ with $\mathcal{R}^c_{[i]}(\Gamma) \subseteq \overline{\{|\psi|\}}$, and $\mathcal{R}_{\Box}(\Delta) \subseteq \overline{\{|\varphi|\}} \cup \{|\psi|\}$. We show that $\otimes_i\psi \in \Delta$. Since $\mathcal{R}^c_{[i]}(\Gamma) \subseteq \overline{\{|\psi|\}}$, there is a $\Sigma \in \mathcal{R}^c_{\Box}(\Delta)$ such that $\neg\psi \in \Sigma$. Hence, $\Diamond\neg\psi \in \Delta$ and so $\neg\Box\psi \in \Delta$. Last, since $\mathcal{R}^c_{\Box}(\Delta) \subseteq \overline{\{|\varphi|\}} \cup \{|\psi|\}$, we have for all $\Gamma \in \mathcal{R}^c_{\Box}(\Delta)$, $\varphi \to \psi \in \Gamma$ (this can be straightforwardly proven by a reductio ad absurdum). Hence, $\Box(\varphi \to \psi) \in \Delta$. Therefore, $\Box(\varphi \to \psi) \land \neg\Box\psi \land \otimes_i\varphi \in \Delta$ and by A22 we have $\otimes_i\psi \in \Delta$.

Since we are using a different kind of canonical model, the modularity of our approach in Section 3.4 does not extend to the present proof. Consequently, we must prove that the properties **Oi** with $i \in \{2, \ldots, 8, 12\}$ also hold for $\mathfrak{M}^{c,+}$ whenever the corresponding axiom is in $\mathsf{OS}_n\mathsf{X}$ extended with A22 (see Table 3.2). We prove that $\mathfrak{M}^{c,+}$ is an $\mathsf{OS}_n\mathsf{X}$-model extended with **O12**. We only consider the more involved cases **O2** and **O8**, the cases **Oi** with $i \in \{3, \ldots, 7\}$ are similar to **O8**. First, we demonstrate **O12**:

O12 Assume $Z \in \mathcal{N}^{c,+}_{\otimes_i}(\Delta)$, there is a $\Sigma \in \mathcal{R}^c_{\Box}(\Delta)$ such that $\mathcal{R}^c_{[i]}(\Sigma) \subseteq \overline{X}$, and (i) $\mathcal{R}^c_{\Box}(\Delta) \subseteq \overline{Z} \cup X$. We show that $X \in \mathcal{N}^{c,+}_{\otimes_i}(\Delta)$. It suffices to consider the case where $Z \notin \mathcal{N}^{c}_{\otimes_i}(\Delta)$ (notice that $\mathcal{N}^{c}_{\otimes_i}$ is the non-supplemented function). Hence, there is a $\otimes_i\varphi \in \Delta$ such that there is a $\Gamma \in \mathcal{R}^c_{\Box}(\Delta)$ with $\mathcal{R}^c_{[i]}(\Gamma) \subseteq \overline{Z}$, and (ii) $\mathcal{R}^c_{\Box}(\Delta) \subseteq \overline{\{|\varphi|\}} \cup Z$. If we show that $\mathcal{R}^c_{\Box}(\Delta) \subseteq \overline{\{|\varphi|\}} \cup X$, then by definition of $\mathcal{N}^{c,+}_{\otimes_i}(\Delta)$ we are done. Suppose towards a contradiction that (iii) $\mathcal{R}^c_{\Box}(\Delta) \not\subseteq \overline{\{|\varphi|\}} \cup X$. Hence, there is a $\Sigma \in \mathcal{R}^c_{\Box}(\Delta)$ such that $\Sigma \notin \overline{\{|\varphi|\}} \cup X$ which means $\Sigma \in \{|\varphi|\} \cap \overline{X}$. By (ii), we know $\Sigma \in Z$, and thus by (i), we have $\Sigma \in X$. Which is a contradiction with (iii).

O2 Assume $Z \in \mathcal{N}^{c,+}_{\otimes_i}(\Delta)$ and suppose towards a contradiction that $\emptyset \in \mathcal{N}^{c,+}_{\otimes_i}(\Delta)$. Two options occur: i) $\emptyset \in \mathcal{N}^{c}_{\otimes_i}(\Delta)$ or ii) $\emptyset \notin \mathcal{N}^{c}_{\otimes_i}(\Delta)$.

i) By definition $\emptyset = \{|\bot|\}$, hence $\otimes_i\bot \in \Delta$ and by A11 $\neg\otimes_i\bot \in \Delta$. Contradiction.

ii) There is a $\otimes_i\varphi \in \Delta$ such that there is a $\Gamma \in \mathcal{R}^c_{\Box}(\Delta)$ with $\mathcal{R}^c_{[i]}(\Gamma) \subseteq \overline{\emptyset} = W$, and $\mathcal{R}^c_{\Box}(\Delta) \subseteq \overline{\{|\varphi|\}}$. Hence, $\Box\neg\varphi \in \Delta$ which gives us $\Box(\varphi \to \bot) \in \Delta$. It is straightforward to show that $\neg\Box\bot \in \Delta$ too. So, $\otimes_i\varphi \land \neg\Box\bot \land \Box(\varphi \to \bot) \in \Delta$ and by A22, $\otimes_i\bot \in \Delta$. By $\otimes_i\varphi$ and A11 we have $\neg\otimes_i\bot \in \Delta$. Contradiction.

O8 Assume $Z \in \mathcal{N}^{c,+}_{\otimes_i}(\Delta)$. We show that (i) there is a $\Gamma \in \mathcal{R}^c_{\Box}(\Delta)$ such that $\mathcal{R}^c_{[i]}(\Gamma) \subseteq Z$, and (ii) there is a $\Sigma \in \mathcal{R}^c_{\Box}(\Delta)$ such that $\mathcal{R}^c_{[i]}(\Sigma) \subseteq \overline{Z}$. It suffices to consider the case where $Z \notin \mathcal{N}^{c}_{\otimes_i}(\Delta)$ ($\mathcal{N}^{c}_{\otimes_i}$ is the non-supplemented function). Hence, there is a $\otimes_i\varphi \in \Delta$ such that (a) there is a $\Omega \in \mathcal{R}^c_{\Box}(\Delta)$ with $\mathcal{R}^c_{[i]}(\Omega) \subseteq \overline{Z}$, and (b) $\mathcal{R}^c_{\Box}(\Delta) \subseteq \overline{\{\!|\varphi|\!\}} \cup Z$.

i) Suppose towards a contradiction that for all $\Gamma \in \mathcal{R}^c_{\Box}(\Delta)$, $\mathcal{R}_{[i]}(\Gamma) \not\subseteq Z$. Hence, for all $\Gamma \in \mathcal{R}^c_{\Box}(\Delta)$ there is a $\Sigma \in \mathcal{R}^c_{[i]}(\Gamma)$, such that $\Sigma \in \overline{Z}$. By (b) it must be that $\Sigma \in \overline{\{\!|\varphi|\!\}}$ and so $\Box\neg[i]\varphi \in \Delta$. However, since $\otimes_i\varphi \in \Delta$, we know by A17 that $\Diamond[i]\varphi \in \Delta$ and so contradiction.

ii) By (a). QED

In adopting A22, the logics for OiLP and OiRz become equivalent, i.e., $\mathsf{OS}_n\{\mathsf{A22},\mathsf{A11}\} \equiv \mathsf{OS}_n\{\mathsf{A22},\mathsf{A12}\}$: For the left-to-right direction, it suffices to show that $\otimes_i\varphi \to \Diamond\varphi$ is a valid formula of $\mathsf{OS}_n\{\mathsf{A22},\mathsf{A11}\}$. Suppose not, then there is an $\mathfrak{M}$ and a $w \in W$ of $\mathfrak{M}$, such that $\mathfrak{M}, w \models \otimes_i\varphi \wedge \neg\Diamond\varphi$. Hence, $\mathfrak{M}, w \models \otimes_i\varphi \wedge \Box\neg\varphi$ and, consequently, $\mathfrak{M}, w \models \Box(\varphi \to \bot)$. By the normality of $\Box$, we have $\mathfrak{M}, w \models \neg\Box\bot$. By an application of axiom A22, we obtain $\mathfrak{M}, w \models \otimes_i\bot$ which is in contradiction with the OiLP implied $\mathfrak{M}, w \models \neg\otimes_i\varphi$. The direction from right-to-left follows directly from the logical taxonomy of OiC (Figure 3.1).

Theorem 3.4 demonstrates that certain $\mathsf{OS}_n\mathsf{X}$ logics can be extended with the restricted monotonicity principle expressed by **O12**. We emphasize that the aim of this extension is to demonstrate that and how it is possible to restore certain intuitive forms of reasoning in the current non-normal modal setting. Although similar extensions are possible of the logics introduced further down below, this is not the aim of the present section. We leave such extensions for some future occasion.

3.5.2 Deontic Contingency

The principle of "Deontic Contingency" (DCg) restricts commands to contingent states of affairs. This means that both what is obligatory and its complement are realizable. The principle implies that neither tautologies nor contradictions can occur within the scope of a deontic operator.[16] Anderson and Moore (1957) discuss DCg in the context of sanctions, requiring that sanctions—i.e., consequences of norm violations—must be both provokable and avoidable. They define deontic concepts in terms of states of affairs implying sanctions (see Chapter 4 for a thorough discussion). We refer to the work of Pascucci (2017) for an extensive formal discussion of the deontic contingency principle. DCg creates room for agency: it ensures that what is obligatory can be potentially influenced by the agents' behavior. Still, the principle allows for contingent states of affairs—such as the occurrence of a moon eclipse—that conceptually lie beyond the grasp of individual agency.

In this respect, DCg captures the same gist as the 'ought implies opportunity' OiO axiom A16. The logical taxonomy of OiC principles (Figure 3.1) tells us that DCg is also satisfied

[16] Von Wright's seminal paper "Deontic Logic" (1951) contains another principle of Deontic Contingency. In Section 3.6, we discuss why his formulation falls short of characterizing contingency.

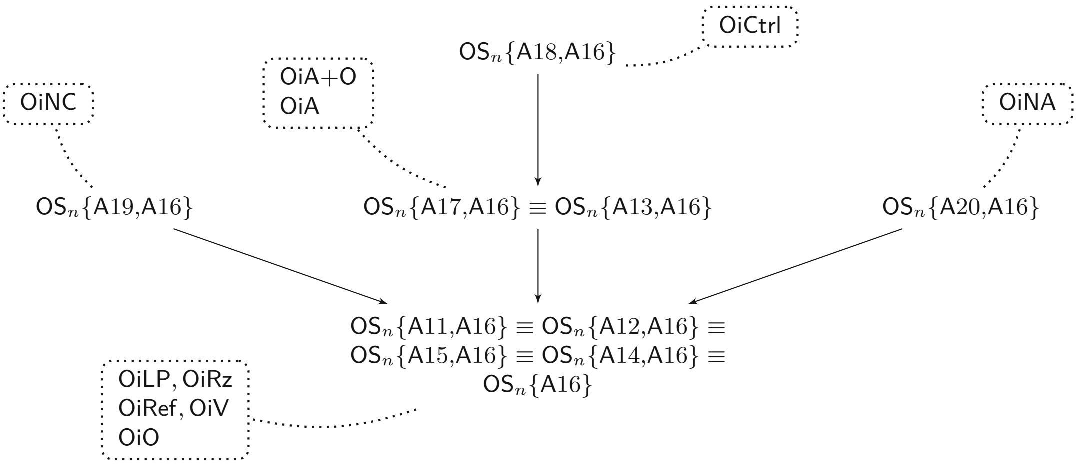

OiCtrl
OS$_n${A18,A16}
OiA+O
OiA
OiNC
OiNA
OS$_n${A19,A16}
OS$_n${A17,A16} ≡ OS$_n${A13,A16}
OS$_n${A20,A16}
OS$_n${A11,A16} ≡ OS$_n${A12,A16} ≡
OS$_n${A15,A16} ≡ OS$_n${A14,A16} ≡
OS$_n${A16}
OiLP, OiRz
OiRef, OiV
OiO

Figure 3.2: Lattice of the ten OiC logics additionally satisfying DCg, i.e., axiom A16.

by OiA+O and OiCtrl. We find that if we adopt DCg as a minimal requirement of deontic logic, several readings of OiC become equivalent. This is expressed in Figure 3.2. That is, enforcing DCg on the ten OiC logics of Figure 3.1 results in five distinct OiC readings as represented in Figure 3.2. The proofs are left out.

3.5.3 Deontic Consistency

The principle of "Deontic Consistency" (DCs) fulfills a central role throughout the history of deontic logic. It requires that obligations are consistent. In a normal modal logic setting, DCs is expressed through the 'the D-axiom' $\neg(\otimes_i \varphi \wedge \otimes_i \neg\varphi)$ (Hilpinen and McNamara, 2013). However, DCs can be interpreted in two distinct ways: first, it can mean that any *single* obligation cannot oblige what is inconsistent, and second, it can mean that any combination of obligations cannot *jointly* oblige what is inconsistent. The two corresponding axioms are, respectively:

A23. $\neg \otimes_i \bot$

A24. $\neg(\otimes_i \varphi \wedge \otimes_i \neg\varphi)$

In normal deontic STIT logics such as DS$_n$ of Chapter 2, these two axioms are equivalent, i.e., $\vdash_{\mathsf{DS}_n} \neg \otimes_i \bot \equiv \neg(\otimes_i \varphi \wedge \otimes_i \neg\varphi)$. This is due to the normality of the $\otimes_i$ operator (the proof is straightforward). However, the move to non-normal modal logics makes it possible to distinguish between "deontic consistency" and the principle of "no deontic dilemmas". In fact, Chellas (1980) takes the conceptual distinction between the two as a reason for adopting non-normal deontic logics. In non-normal deontic logics, $\neg \otimes_i \bot$ ensures that no obligation is inconsistent, whereas $\neg(\otimes_i \varphi \wedge \otimes_i \neg\varphi)$ ensures that there are no deontic dilemmas. Hence, only A23 expresses DCs proper. The principle of "No Deontic Dilemmas" is discussed extensively in the next section.

All $\mathsf{OS}_n\mathsf{X}$ logics from Figure 3.1 that imply OiRz also imply $\neg\otimes_i\bot$ ($\mathsf{A23}$). The logic $\mathsf{OS}_n\{\mathsf{A11}\}$ for OiLP is strictly subsumed by the logic $\mathsf{OS}_n\{\mathsf{A23}\}$. The proof is straightforward. Hence, the logic $\mathsf{OS}_n\{\mathsf{A23}\}$ is located in between $\mathsf{OS}_n\{\mathsf{A12}\}$ and $\mathsf{OS}_n\{\mathsf{A11}\}$ in Figure 3.1. Still, we would argue that OiLP captures the same gist as DCs since it expresses that either there is some obligation $\otimes_i\varphi$, and so $\neg\otimes_i\bot$ is implied, or there are no obligations at all, and so *a fortiori* no inconsistent obligations either.

The frame property characterizing axiom $\mathsf{A23}$ is:

O13 For all $w \in W, \emptyset \notin \mathcal{N}_{\otimes_i}(w)$

All considered logics can be extended with this axiom. (Recall that in some cases, the axiomatization is not minimal since $\mathsf{A23}$ is already implied.) The lattice of Figure 3.1 is preserved with the exception of additional arrows from $\mathsf{OS}_n\{\mathsf{A14{,}A23}\}$, $\mathsf{OS}_n\{\mathsf{A19{,}A23}\}$, and $\mathsf{OS}_n\{\mathsf{A20{,}A23}\}$ to $\mathsf{OS}_n\{\mathsf{A11{,}A23}\}$. The resulting logics are sound and complete.

Theorem 3.5. *Any $\mathsf{OS}_n\mathsf{X}$ logic extended with axiom $\mathsf{A23}$ is sound and complete with respect to their corresponding class of $\mathsf{OS}_n\mathsf{X}$-frames extended with **O13**.*

Proof. It suffices to only consider the additional case for $\mathsf{A23}$ and **O13**.

Soundness. Take an arbitrary $\mathsf{OS}_n\mathsf{X}$-model $\mathfrak{M}$ extended with **O13** and let $w \in W$. Since $\emptyset \notin \mathcal{N}_{\otimes_i}(w)$ and $\emptyset = \|\bot\|$, by semantic definition of $\otimes_i$ we have $\mathfrak{M}, w \models \neg\otimes_i\bot$.

Completeness. Let $\mathfrak{M}^c$ be a canonical model for the logic $\mathsf{OS}_n\mathsf{X}$ extended with axiom $\mathsf{A23}$. Let $\Delta \in W^c$. By duality, $\neg\otimes_i\bot$ is equivalent to $\ominus_i\top$. By Lemma 3.6 and the fact that $\ominus_i\top \in \Delta$, we have $\overline{\{\!|\top|\!\}} \notin \mathcal{N}^c_{\otimes_i}(\Delta)$. Since $\{\!|\top|\!\} = W$, we have $\overline{\{\!|\top|\!\}} = \emptyset \notin \mathcal{N}^c_{\otimes_i}(\Delta)$. QED

3.5.4 No Deontic Dilemmas

A *deontic dilemma* is a situation in which an agent “both ought to do something and ought not to do that thing” (Lemmon, 1962, p.148). So far, all the developed $\mathsf{OS}_n\mathsf{X}$ logics can consistently model deontic dilemmas. In other words, each $\mathsf{OS}_n\mathsf{X}$ logic is compatible with the *existence* of deontic dilemmas.

The principle of “No Deontic Dilemmas” (NDD) stipulates that no two obligations can prescribe jointly inconsistent state of affairs. Its corresponding axiom is:

$\mathsf{A24}$. $\neg(\otimes_i\varphi \wedge \otimes_i\neg\varphi)$

The corresponding frame property for this principle is defined as follows:

O14 For all $w \in W, Z \subseteq W$, if $Z \in \mathcal{N}_{\otimes_i}(w)$, then $\overline{Z} \notin \mathcal{N}_{\otimes_i}(w)$

The resulting logics are sound and complete.

Theorem 3.6. *Any* $\mathsf{OS}_n\mathsf{X}$ *logic extended with axiom* $\mathsf{A24}$ *is sound and complete with respect to their corresponding class of* $\mathsf{OS}_n\mathsf{X}$*-frames extended with* ***O14****.*

Proof. It suffices to only consider the additional case for $\mathsf{A24}$ and **O14**.

Soundness. Take an arbitrary $\mathsf{OS}_n\mathsf{X}$-model $\mathfrak{M}$ extended with **O14** and let $w \in W$. Observe that $\models_{\mathsf{OS}_n\mathsf{X}} \neg(\otimes_i\varphi \wedge \otimes_i\neg\varphi) \equiv \otimes_i\varphi \rightarrow \neg\otimes_i\neg\varphi$. Assume $\mathfrak{M}, w \models \otimes_i\varphi$, then $\|\varphi\| \in \mathcal{N}_{\otimes_i}(w)$ and by **O14**, $\overline{\|\varphi\|} \notin \mathcal{N}_{\otimes_i}(w)$. Since $\overline{\|\varphi\|} = \|\neg\varphi\|$ by Lemma 3.1-(v), we have $\|\neg\varphi\| \notin \mathcal{N}_{\otimes_i}(w)$ and so $\mathfrak{M}, w \models \neg \otimes_i \neg\varphi$.

Completeness. Let $\mathfrak{M}^c$ be a canonical model for the logic $\mathsf{OS}_n\mathsf{X}$ extended with the axiom $\mathsf{A23}$. Let $\Delta \in W^c$. Suppose $Z \in \mathcal{N}^c_{\otimes_i}(\Delta)$, hence there is an $\otimes_i\varphi \in \Delta$ such that $Z = \{\!|\varphi|\!\}$. Since $\otimes_i\varphi \rightarrow \neg \otimes_i \neg\varphi \in \Delta$, we have $\neg\otimes_i\neg\varphi \in \Delta$ and by Lemma 3.6, we have $\overline{\{\!|\varphi|\!\}} \notin \mathcal{N}^c_{\otimes_i}(\Delta)$. QED

Interestingly, adding NDD to $\mathsf{OS}_n\mathsf{X}$ does not change the logical taxonomy of OiC in Figure 3.1. In other words, the metaethical principle of "No Deontic Dilemmas" is *logically independent* from OiC. This means that the existence of moral dilemmas does not necessarily entail the rejection of OiC. In fact, of those philosophers endorsing the existence of moral dilemmas, some reject and some accept OiC (see (Marcus, 1980) for an overview). We now discuss Lemmon's rejection of OiC.

Reasoning with Dilemmas and OiC: Aggregation. NDD is a minimal principle for most deontic logics (Hilpinen and McNamara, 2013). However, some accounts refute NDD as a basic principle of deontic logic. Most notably, Lemmon (1962) advocates the existence of moral dilemmas: "It is a nasty fact about human life that we sometimes both ought and ought not to do things; but it is not a logical contradiction" (p.150). Consequently, he argues that if "ought implies can" holds, a contradiction is obtainable from a moral dilemma. According to Lemmon, a deontic dilemma between $\otimes_i\varphi$ and $\otimes_i\neg\varphi$ logically implies the inconsistent obligation $\otimes_i(\varphi\wedge\neg\varphi)$ and since $\varphi\wedge\neg\varphi$ is impossible—i.e., $\neg\Diamond(\varphi \wedge \neg\varphi)$—an adaptation of $\otimes_i\varphi \rightarrow \Diamond\varphi$ (OiRz) would imply that $\otimes_i\varphi$ and $\otimes_i\neg\varphi$ are logically inconsistent. And so, Lemmon concludes, "I view this [...] as a refutation of the principle that 'ought' implies 'can' " (Lemmon, 1962, p.150). Let us look at this argument in more detail.

The logic considered by Lemmon is a *normal* modal deontic logic. The inconsistency is a consequence of *two* deontic reasoning principles interacting: the aggregation of deontic modalities (i.e., C) together with the principle of OiC. The analysis provided in this chapter demonstrates that OiC is, in fact, logically compatible with deontic dilemmas, even in the light of the principle of Deontic Consistency $\neg \otimes_i \bot$. However, the aggregation principle C is incompatible with deontic dilemmas in light of Deontic

Consistency. Therefore, we must conclude that inconsistency as a result of the formal representation of deontic dilemmas is more a problem of aggregation than of OiC.[17]

A common approach is to loose the aggregation principle C, e.g., (Chellas, 1980). However, as remarked on page 78, the logic may become too weak for logical reasoning with obligations.[18] Fortunately, not all is lost: it remains possible to adopt forms of *restricted* aggregation, which only allow for the aggregation of jointly consistent obligations. In the context of OiRz, we can adopt the following axiom:

$\mathsf{A25}$. $(\Diamond(\varphi \wedge \psi) \wedge \otimes_i\varphi \wedge \otimes_i\psi) \rightarrow \otimes_i(\varphi \wedge \psi)$

Axiom $\mathsf{A25}$ enables the aggregation of two obligations, provided they are jointly realizable. Such an aggregation principle should ideally take into account the reading of OiC adopted in the logics to which the principle is added. For instance, in light of OiA, the previous axiom may be considered insufficient. In that context, if $\otimes_i\varphi$ and $\otimes_i\psi$ hold, then agent i is only under the obligation of φ and ψ together, whenever i has the *ability* to see to it that both φ and ψ hold. In that case, we can adopt the following axiom:

$\mathsf{A26}$. $(\Diamond[i](\varphi \wedge \psi) \wedge \otimes_i\varphi \wedge \otimes_i\psi) \rightarrow \otimes_i(\varphi \wedge \psi)$

In other words, we can reintroduce deontic reasoning principles to $\mathsf{OS}_n\mathsf{X}$ by taking into account specific readings of OiC. It is also possible to add unrestricted C to all logics. However, the latter principle is arguably too strong since it excludes the possibility of dilemmas altogether.

The corresponding frame properties are,

O15 For all $w \in W, Z, X \subseteq W$, if $Z, X \in \mathcal{N}_{\otimes_i}(w)$ and $\mathcal{R}_\Box(w) \cap Z \cap \mathcal{X} \neq \emptyset$, then $Z \cap X \in \mathcal{N}_{\otimes_i}(w)$

respectively,

O16 For all $w \in W, Z, X \subseteq W$, if $Z, X \in \mathcal{N}_{\otimes_i}(w)$ and there is a $v \in W$, such that $\mathcal{R}_{[i]}(v) \subseteq \mathcal{R}_\Box(w) \cap Z \cap X$, then $Z \cap X \in \mathcal{N}_{\otimes_i}(w)$

All $\mathsf{OS}_n\mathsf{X}$ logics extended with these versions of restricted aggregation are sound and complete.

[17] Still, Lemmon (1962) remarks that "there are surely clear counterexamples [to OiC] even without the introduction of the present instances [dilemmas]" (p.150). Such examples are not given in that paper, but we refer to the work of Vranas (2007) and Vranas (2018a) for several objections to OiC.

[18] Van Fraassen's (1973) account is arguably the first attempt to weaken deontic logic in order to accommodate deontic dilemmas. See the work of Horty (1994) for an alternative approach. The aim is to provide a formalism that blocks deductive explosion from deontic dilemmas (such as in normal modal deontic logics that contain a D-axiom) while maximizing inferential power. See also the discussion of the Alternative Service Paradox on Section 1.2.1.

Theorem 3.7. *Any* $\mathsf{OS}_n\mathsf{X}$ *logic extended with axiom* $\mathsf{A25}$*, respectively axiom* $\mathsf{A26}$ *is sound and complete with respect to its corresponding class of* $\mathsf{OS}_n\mathsf{X}$*-frames extended with* ***O15****, respectively* ***O16****.*

Proof. It suffices to only consider the additional case for $\mathsf{A25}$ and **O15**. The proof of $\mathsf{A26}$ and **O16** is similar.

Soundness. Take an arbitrary $\mathsf{OS}_n\mathsf{X}$-model $\mathfrak{M}$ extended with **O15** and let $w \in W$. Assume $\mathfrak{M}, w \models \Diamond(\varphi \wedge \psi) \wedge \otimes_i\varphi \wedge \otimes_i\psi$, then by semantic definitions of $\otimes_i$ and $\Box$, we know that $\|\varphi\|, \|\psi\| \in \mathcal{N}_{\otimes_i}(w)$ and there is a $v \in \mathcal{R}_\Box(w)$ such that $\mathfrak{M}, v \models \varphi \wedge \psi$, and so $\mathcal{R}_\Box(w) \cap \|\varphi\| \cap \|\psi\| \neq \emptyset$. By **O15**, $\|\varphi\| \cap \|\psi\| \in \mathcal{N}_{\otimes_i}(w)$, and so $\mathfrak{M}, w \models \otimes_i(\varphi \wedge \psi)$.

Completeness. Let $\mathfrak{M}^c$ be a canonical model for the logic $\mathsf{OS}_n\mathsf{X}$ extended with the axiom $\mathsf{A25}$. Let $\Delta \in W^c$. Suppose $Z, X \in \mathcal{N}^c_{\otimes_i}(\Delta)$ and $\mathcal{R}^c_\Box(\Delta) \cap Z \cap X \neq \emptyset$. Hence, there are $\otimes_i\varphi, \otimes_i\psi \in \Delta$ such that $Z = \{\!|\varphi|\!\}$ and $X = \{\!|\psi|\!\}$, by construction of $\mathfrak{M}^c$. Since $\mathcal{R}^c_\Box(\Delta) \cap \{\!|\varphi|\!\} \cap \{\!|\psi|\!\} \neq \emptyset$, there is a $\Gamma \in \mathcal{R}^c_\Box(\Delta)$ such that $\Gamma \in \{\!|\varphi \wedge \psi|\!\}$ and so, by Lemma 3.7, $\Diamond(\varphi \wedge \psi) \in \Delta$ too. Hence, since $\Diamond(\varphi \wedge \psi) \wedge \otimes_i\varphi \wedge \otimes_i\psi \rightarrow \otimes_i(\varphi \wedge \psi) \in \Delta$, we have $\otimes_i(\varphi \wedge \psi) \in \Delta$, and therefore $Z \cap X \in \mathcal{N}_{\otimes_i}(\Delta)$. QED

Reasoning with Dilemmas and OiC: Disjunction. As a final remark on deontic dilemmas, we discuss the notion of *disjunctive response*. A common answer to deontic dilemmas is that the agent is at least under the obligation to choose. To choose can be seen as the lesser of two evils. Namely, although it is impossible to comply with both obligations, complying with one is better than not complying.[19] The idea of disjunctive response is captured through the axiom,

$\mathsf{A27}$. $(\otimes_i\varphi \wedge \otimes_i\neg\varphi) \rightarrow \otimes_i(\varphi \vee \neg\varphi)$

and the frame property

O17 For all $w \in W$, $Z \subseteq W$, if $Z, \overline{Z} \in \mathcal{N}_{\otimes_i}(w)$, then $Z \cup \overline{Z} \in \mathcal{N}_{\otimes_i}(w)$

Two remarks are in place here. First, in a non-normal modal logic such as OS_n, the formula (†) $\otimes_i\varphi \rightarrow \otimes_i(\varphi \vee \psi)$ is not a theorem due to the absence of the property of monotonicity M. Consequently, Ross' paradox—i.e., if one ought to post a letter, one ought to either post it or burn it—does not necessarily hold in a non-normal modal setting (Chapter 1). The above axiom, however, is different from (†) since it only introduces a disjunction from two conflicting obligations.

Theorem 3.8. *Any* $\mathsf{OS}_n\mathsf{X}$ *logic extended with axiom* $\mathsf{A27}$ *is sound and complete with respect to its corresponding class of* $\mathsf{OS}_n\mathsf{X}$*-frames extended with* ***O17****.*

[19] See Chapter 5 for a discussion of similar principle called *Vikalpa*, adopted by the ancient South Asian school called Mīmāṃsā.

Proof. It suffices to only consider the additional case for A27 and **O17**.

Soundness. Take an arbitrary $\mathsf{OS}_n\mathsf{X}$-model $\mathfrak{M}$ extended with **O17** and let $w \in W$. Assume $\mathfrak{M}, w \models \otimes_i \varphi \wedge \otimes_i \neg\varphi$. By semantic definition $\|\varphi\|$,$\|\neg\varphi\| \in \mathcal{N}_{\otimes_i}(w)$, be the definition of truth set, $\overline{\|\varphi\|} \in \mathcal{N}_{\otimes_i}(w)$. By **O17**, $\|\varphi\| \cup \overline{\|\varphi\|} \in \mathcal{N}_{\otimes_i}(w)$ and so $\mathfrak{M}, w \models \otimes_i(\varphi \vee \neg\varphi)$.

Completeness. Let $\mathfrak{M}^c$ be a canonical model for the logic $\mathsf{OS}_n\mathsf{X}$ extended with the axiom A27. Let $\Delta \in W^c$. Assume $Z, \overline{Z} \in \mathcal{N}^c_{\otimes_i}(\Delta)$. Then by construction of $\mathfrak{M}^c$, there are $\otimes_i\varphi, \otimes_i\neg\varphi \in \Delta$ such that $Z = \{|\varphi|\}$. By $(\otimes_i\varphi \wedge \otimes_i\neg\varphi) \rightarrow \otimes_i(\varphi \vee \neg\varphi) \in \Delta$, we have $\otimes_i(\varphi \vee \neg\varphi) \in \Delta$ and so $Z \cup \overline{Z} \in \mathcal{N}^c_{\otimes_i}(\Delta)$ too. QED

Remark 3.5. *In the light of* NVC*, adding axiom* A27 *excludes the possibility of deontic dilemmas. Namely, since* NVC *is expressed as* $\neg \otimes_i \top$*, by contraposition on* A27 *we have* $\neg \otimes_i \top \rightarrow \neg(\otimes_i\varphi \wedge \otimes_i\neg\varphi)$*, and so we obtain* $\neg(\otimes_i\varphi \wedge \otimes_i\neg\varphi)$*. In other words, all deliberative readings of OiC implying* NVC *imply the endorsement of* NDD*.*

A more restricted version of A27 is the following axiom:

A28. $\Box([i]\varphi \rightarrow \neg[i]\psi) \wedge \otimes_i\varphi \wedge \otimes_i\psi \wedge (\neg\Box(\varphi \vee \psi)) \rightarrow \otimes_i(\varphi \vee \psi)$

The axiom expresses the following: if the agent is subject to two obligations that cannot be jointly seen to, then she is obliged to satisfy at least one of the two obligations, *provided* it is not trivial that either of the two obligations is satisfied. The last conjunct of the antecedent of A28 ensures that introducing a deontic disjunction does not imply a tautological obligation. We stress that extending an $\mathsf{OS}_n\mathsf{X}$ logic with A28 does preserve the possibility of consistently modeling deontic dilemmas, even in the context of NVC.

The corresponding frame property is as follows:

O18 For all $w \in W, Z, X \subseteq W$, if $\mathcal{R}_\Box(w) \not\subseteq Z \cup X, Z, X \in \mathcal{N}_{\otimes_i}(w)$, and for all $v \in \mathcal{R}_\Box(w)$, $\mathcal{R}_{[i]}(v) \subseteq X$, implies $\mathcal{R}_{[i]}(v) \not\subseteq X$, then $Z \cup X \in \mathcal{N}_{\otimes_i}(w)$

The resulting logics are sound and complete.

Theorem 3.9. *Any* $\mathsf{OS}_n\mathsf{X}$ *logic extended with axiom* A28 *is sound and complete with respect to its corresponding class of* $\mathsf{OS}_n\mathsf{X}$*-frames extended with* ***O18****.*

Proof. It suffices to only consider the additional case for A28 and **O18**.

Soundness. Take an arbitrary $\mathsf{OS}_n\mathsf{X}$-model $\mathfrak{M}$ extended with **O18** and let $w \in W$. Assume $\mathfrak{M}, w \models \Box([i]\varphi \rightarrow \neg[i]\psi) \wedge \otimes_i\varphi \wedge \otimes_i\psi \wedge \neg\Box(\varphi \vee \psi)$. By the semantic definitions of $\otimes_i, \Box$, and $[i]$ we know that $\|\varphi\|, \|\psi\| \in \mathcal{N}_{\otimes_i}(w)$, $\mathcal{R}_\Box(w) \not\subseteq \|\varphi \vee \psi\|$, and for all $v \in \mathcal{R}_\Box(w)$, $\mathcal{R}_{[i]}(v) \subseteq \|\varphi\|$ implies $\mathcal{R}_{[i]}(v) \not\subseteq \|\psi\|$. By **O18** we have $\|\varphi\| \cup \|\psi\| \in \mathcal{N}_{\otimes_i}(w)$ and so $\mathfrak{M}, w \models \otimes_i(\varphi \vee \psi)$.

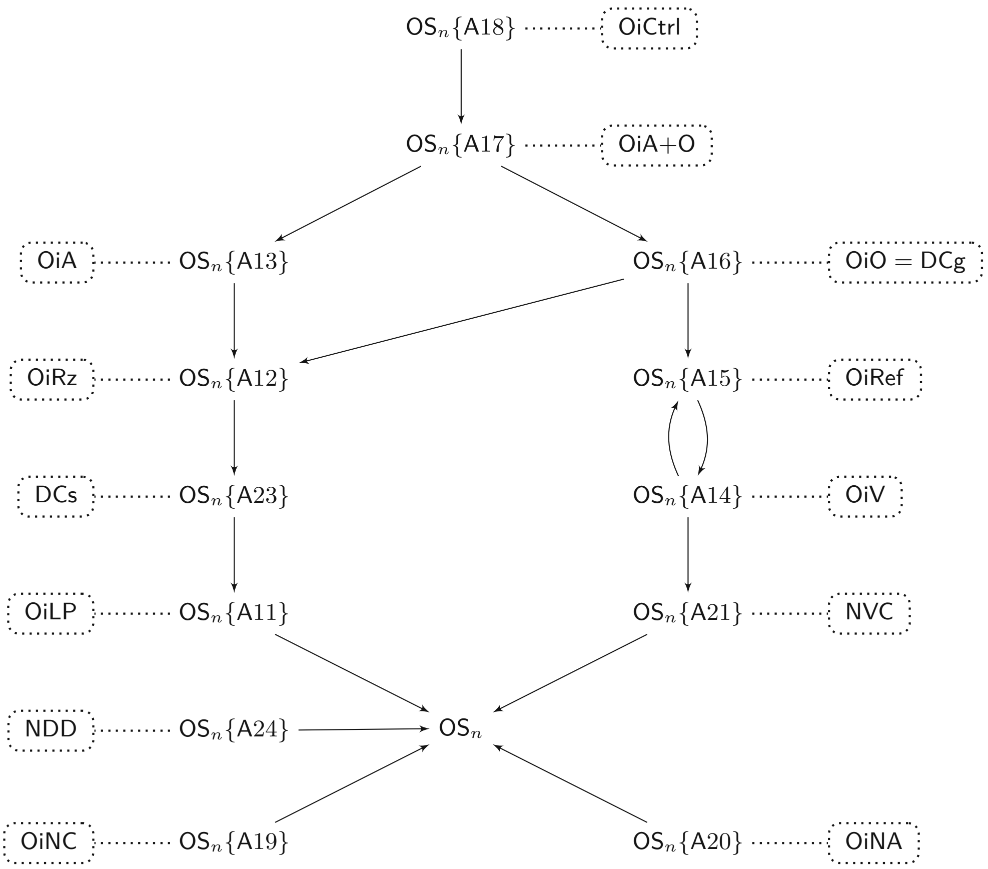


Figure 3.3: The lattice of metaethical principles NVC, DCg, DCs, and NDD, together with the ten OiC principles of Table 3.1. Directed edges point from subsuming to subsumed logics, e.g., $\mathsf{OS}_n \subseteq \mathsf{OS}_n\{\mathsf{A18}\}$. Reflexive and transitive edges are omitted. Dotted nodes make explicit which logics correspond to which metaethical principle.

Completeness. Let $\mathfrak{M}^c$ be a canonical model for the logic $\mathsf{OS}_n\mathsf{X}$ extended with the axiom A25. Let $\Delta \in W^c$. Assume $\mathcal{R}^c_{\Box}(\Delta) \not\subseteq Z \cup X$, $Z, X \in \mathcal{N}^c_{\otimes_i}(\Delta)$, and for all $\Gamma \in \mathcal{R}^c_{\Box}(\Delta), \mathcal{R}^c_{[i]}(\Gamma) \subseteq Z$ implies $\mathcal{R}^c_{[i]}(\Gamma) \not\subseteq X$. By the construction of $\mathfrak{M}^c$, there are $\otimes_i\varphi, \otimes_i\psi \in \Delta$ such that $\{\!|\varphi|\!\} = Z$ and $\{\!|\psi|\!\} = X$. Since $\mathcal{R}^c_{\Box}(\Delta) \not\subseteq \{\!|\varphi|\!\} \cup \{\!|\psi|\!\}$, there is a $\Gamma \in \mathcal{R}^c_{\Box}(\Delta)$ such that $\neg(\varphi \vee \psi) \in \Gamma$. By Lemma 3.7 and the semantic definition of $\Box$, $\neg\Box(\varphi \vee \psi) \in \Delta$. With similar reasoning, we obtain $\Box([i]\varphi \to \neg[i]\psi) \in \Delta$. Since axiom A28 $\in \Delta$, we know that $\otimes_i(\varphi \vee \psi) \in \Delta$ and so $Z \cup X \in \mathcal{N}^c_{\otimes_i}(\Delta)$. QED

3.5.5 A Logical Taxonomy of Metaethical Principles

In conclusion, we present Figure 3.3, which demonstrates the logical relations between all OiC principles, as well as all the other metaethical principles discussed (Objective 3).

Generalizing the endorsement principle of Definition 3.9, Figure 3.3 demonstrates which endorsed metaethical principles commit one to endorsing other metaethical principles in the context of STIT.

We find that the metaethical principles NVC, DCg, and DCs are logically related to OiC. Namely, they imply or are implied by certain readings of OiC. In particular, DCs is implied by all OiC readings implying OiRz, whereas NVC is implied by all OiC readings implying OiV. This logical dependency is presented in Figure 3.3. Furthermore, all metaethical principles can be consistently added to the class of $\mathsf{OS}_n\mathsf{X}$ logics.

We also showed that the adaptation of metaethical principles, such as NVC and DCg, make certain readings of OiC equivalent in the context of STIT, thus enforcing certain interpretations of OiC. For instance, if one adopts DCg together with OiA, then one is logically committed to endorsing OiA+O (see Figure 3.3).

Furthermore, Figure 3.3 shows that the principle NDD does not logically relate to any OiC interpretations. This is contrary to the claim made by Lemmon (1962) that the existence of moral dilemmas logically refutes OiC. We argued that Lemmon's claim strongly depends on the endorsement of reasoning principle C for $\otimes_i$, which enables the aggregation of (conflicting) obligations. By using non-normal modal logics, we obtain a deeper understanding of the logical interdependencies of various metaethical principles important in ethics and the field of deontic logic.

Last, we illustrated how the class of $\mathsf{OS}_n\mathsf{X}$ logics can be extended with deontic reasoning principles that take into account the different OiC readings (Objective 4). For instance, we provided a restricted monotonicity principle A22 that is compatible with readings of OiC that imply NVC. Furthermore, we extended $\mathsf{OS}_n\mathsf{X}$ with restricted aggregation principles A25 and A26, and a principle modeling disjunctive response A27. All the resulting logics are sound and strongly complete.

3.6 Related Work and Future Research

Vranas and OiC. In a series of papers (Vranas, 2007; Vranas, 2018a; Vranas, 2018b), Vranas discusses and refutes several objections to OiC. Most objections relate to counterfactual reasoning, temporal aspects such as obligations with deadlines, conditional obligations, and notions of culpability, blame, and moral judgment. He argues that, although OiC remains a controversial principle for several reasons, the discussed objections can all be refuted. Vranas (2018b) adopts a temporal reading of OiC. His account deals with pro tanto and all things considered obligations and how certain obligations remain, cease to be, and are overridden in conflicting scenarios. Vranas' (2018a) account of OiC is "if an agent at a given time has an obligation, then the agent at that time can obey the obligation" (p.23) where 'can obey' equals ability plus opportunity (cf. OiA+O). Vranas' (2007; 2018a) account of OiC is inherently defeasible since it deals with agents losing their obligations over time due to agents becoming unable to obey the obligation.

We agree with Vranas that there is a strong connection between OiC, temporal reasoning, and the defeasibility of obligations. One can think of OiC as a restriction on all-things-considered reasoning with obligations. Consider a scenario in which I am obliged to keep a promise until Sunday. In a CTD scenario (Chapter 2), one might have the CTD obligation to apologize if one does not keep the promise, e.g., after Sunday. However, what happens if I know on Saturday already that I cannot keep the promise due to circumstances beyond my control? In some cases, one may argue that OiC entails that my initial obligation to keep the promise ceases to be (Vranas, 2018b). In that sense, being unable is a constraint that makes obligations defeasible. Temporal questions concerning CTD reasoning and OiC are challenging, and the investigation of OiC as a defeasible principle deserves further investigation.

Open question 3.1. *Can we give a formal account of defeasible interpretations of OiC, e.g., where 'can' is considered a normality and 'cannot' an abnormality that causes the revision of the inferred obligations?*

Concerning the above, Vranas (2007) discusses a scenario in which an agent is obliged to submit an essay within a given time interval $t_1, \dots, t_n$, where the agent can fulfill her obligation within this interval. One of the challenges of OiC is determining whether the obligation is annulled if the agent waits until she cannot fulfill her obligation anymore (for instance, when she procrastinates writing the essay until an hour before the deadline t_n). What is the difference between an agent not being able to fulfill her obligations and an agent deliberately seeing to it that she cannot fulfill her obligations? There is an interesting connection here with other agentive concepts such as culpability, blameworthiness, responsibility, and causal contribution. We leave such investigations for future work.

Other OiC Readings. We briefly mention some OiC principles not discussed in this chapter. Prakken and Sergot (1996) adopt an OiC principle similar to OiRz. Their conditional logic contains a principle of the form $\otimes_i^{\psi}\varphi \to \Diamond\varphi$ which means that "if in context ψ, φ is ideal then φ is possible (even though φ might not be possible in context ψ)". The conditional obligation requires 'can' in OiC to be global, i.e., irrespective of the context of the obligation.

Vranas (2018a) proposes several other metaethical principles which generalize OiC. For instance, the principle 'ought-implies-can-obey', 'ought-implies-can-satisfy', 'ought-implies-possible-violation', and 'ought-implies-can-avoid'. See also (Vranas, 2018b) for a discussion of the influence of agents' epistemic limitation on OiC and all things considered obligations. There is some immediate terminological overlap between the terms used in the above principles and the OiC readings discussed in this chapter. However, the exact relations remain to be determined. In particular, Vranas discusses these principles in the context of conditional obligations and temporal agentive reasoning. The logic in this chapter is atemporal. For these reasons, we must postpone a proper formal investigation of the principles defended in (Vranas, 2018a).

Last, Broersen (2003) discusses an 'ought implies may' principle in a modal action logic setting. The principle captures the idea that if some action α is obligatory, it is permitted to perform α, i.e., the agent may perform α. The principle is formally represented as $\mathcal{O}\varphi \rightarrow \mathcal{P}\varphi$. Here, permission ($\mathcal{P}$) and obligation ($\mathcal{O}$) are reduced to statements concerning actions leading to violations; cf. (Meyer, 1988). In Chapter 4, we discuss such reductions in detail. It must be noted that in the 'ought implies may' reading, permission is not definable in terms of obligations, such as in Standard Deontic Logic (Hilpinen and McNamara, 2013) (see page 13).

Other Metaethical Principles. In the seminal work "Deontic Logic" (1951), von Wright argues for the adaptation of, what he calls, the principle of deontic contingency (cf. page 107). He phrases it accordingly "[a] tautologous act is not necessarily obligatory, and a contradictory act is not necessarily forbidden" (p.11). Let $\mathcal{O}$ be an obligation operator and $\mathcal{F}$ a prohibition operator. The principle concerns necessitation and can be expressed as $\nvdash \mathcal{O}\top$ and $\nvdash \mathcal{F}\bot$. In Standard Deontic Logic, the two remarks about obligations and prohibition are equivalent since $\mathcal{F}\varphi = \mathcal{O}\neg\varphi$. Von Wright's proposal amounts to omitting necessitation as a property of the obligation operator. For that reason, von Wright's original Deontic Logic (1951) can be taken as a non-normal modal logic. We point out that although von Wright refers to this principle as a principle of contingency, it is too weak to guarantee contingency. Namely, that tautologous acts are not *necessarily* obligatory leaves room for some obligations to be tautologies and, consequently, not contingent. None of the logics $\mathsf{OS}_n\mathsf{X}$ satisfies necessitation N and, thus, von Wright's principle holds for these logics.

The ancient south Asian school of Mīmāṃsā—devoted to the structural analysis of normative statements in the prescriptive part of the Vedas—proposed various metaethical principles. For instance, for the Mīmāṃsā, actions occurring in commands must be meaningful. An action is not meaningful whenever the agent is naturally inclined to comply in any given case or whenever the prescribed action is impossible to fulfill (van Berkel et al., 2022a; Freschi and Pascucci, 2021). The former is related to NVC, whereas the latter expresses DCs. In Chapter 5, we provide a formalization of the deontic theory of the Mīmāṃsā philosopher Maṇḍana and discuss various related metaethical principles in detail.

Quasi-Agentive Obligations and OiC. Last, we point out that none of the logics presented in this chapter is equivalent to the traditional deontic STIT logic DS_n (Horty, 2001; Murakami, 2005) (cf. Chapter 2). The logic DS_n is a normal modal logic, which requires that the modality $\otimes_i$ satisfies normality, e.g., the axioms M, C, and N (Section 3.2). What is more, in DS_n, the formula

$$\otimes_i\varphi \equiv \otimes_i[i]\varphi$$

is a theorem. It characterizes the *quasi-agentive* reading of the obligation (Belnap and Perloff, 1988), by equating each obligation for an agent with an obligatory *choice* for that agent: "agent i ought to see to it that". In this chapter, we deliberately abstained from adopting the quasi-agentive obligation operator. We now discuss the implications of adopting this operator for the analysis of OiC and pose some open questions.

In order to obtain the quasi-agentive reading of $\otimes_i$ in $\mathsf{OS}_n\mathsf{X}$, we must impose additional properties. We argue that to obtain the quasi-agentive reading of obligation, it is enough to consider $\otimes_i\varphi \rightarrow \otimes_i[i]\varphi$ because it sufficiently ensures that each obligation $\otimes_i\varphi$ has a corresponding quasi-agentive obligation $\otimes_i[i]\varphi$. Guaranteeing the quasi-agentive reading of $\otimes_i$ is not trivial. We can adopt the following axiom:

A29. $\otimes_i\varphi \rightarrow \otimes_i[i]\varphi$

We conjecture that axiom A29 corresponds to the following frame property:

O19 For all $w \in W, Z \subseteq W$, if $Z \in \mathcal{N}_{\otimes_i}(w)$, then $\{v \in W \mid \mathcal{R}_{[i]}(v) \subseteq Z\} \in \mathcal{N}_{\otimes_i}(w)$

It is left for future work to determine whether $\mathsf{OS}_n\mathsf{X}$ extended with axiom A29 is sound and complete with respect to the class of $\mathsf{OS}_n\mathsf{X}$-models extended with **O19**.

Open question 3.2. *Under which conditions can the quasi-agentive reading of the* STIT *obligation* $\otimes_i$ *be restored for non-normal modal* STIT *logics* $\mathsf{OS}_n\mathsf{X}$*?*

In the remainder, we briefly discuss the effects of adding A29 to the axiomatic characterization of $\mathsf{OS}_n\mathsf{X}$. It can be directly observed that certain readings of OiC become equivalent. To illustrate, consider a logic $\mathsf{OS}_n\mathsf{X}$ with $\otimes_i\varphi \rightarrow \Diamond\varphi$ (OiRz) as a theorem. Furthermore, assume $\otimes_i\varphi \rightarrow \otimes_i[i]\varphi$ is an axiom of that system. By straightforward propositional reasoning, we obtain the following theorem:

$$\otimes_i\varphi \rightarrow \Diamond[i]\varphi$$

In other words, adding the quasi-agentive reading of $\otimes_i$ implies that the OiC readings of OiRz and OiA become equivalent. Given axiom A29, one can construct similar arguments that show the equivalence of OiNC and OiNA and of OiO and OiA+O. Consequently, under the quasi-agentive reading of $\otimes_i$, there are strictly fewer variations of OiC. In fact, of the ten principles discussed, at most six readings are preserved. This is expressed in Figure 3.4. We conjecture that the lattice in Figure 3.4 represents the resulting interdependencies of OiC logics strengthened with the quasi-agentive obligation $\otimes_i\varphi \rightarrow \otimes_i[i]\varphi$.

* * *

This chapter provided a comprehensive logical study of the variety of *Ought implies Can* (OiC). We analyzed ten principles from the philosophical literature and provided formalizations of each of them. We developed a class of sound and complete deontic STIT

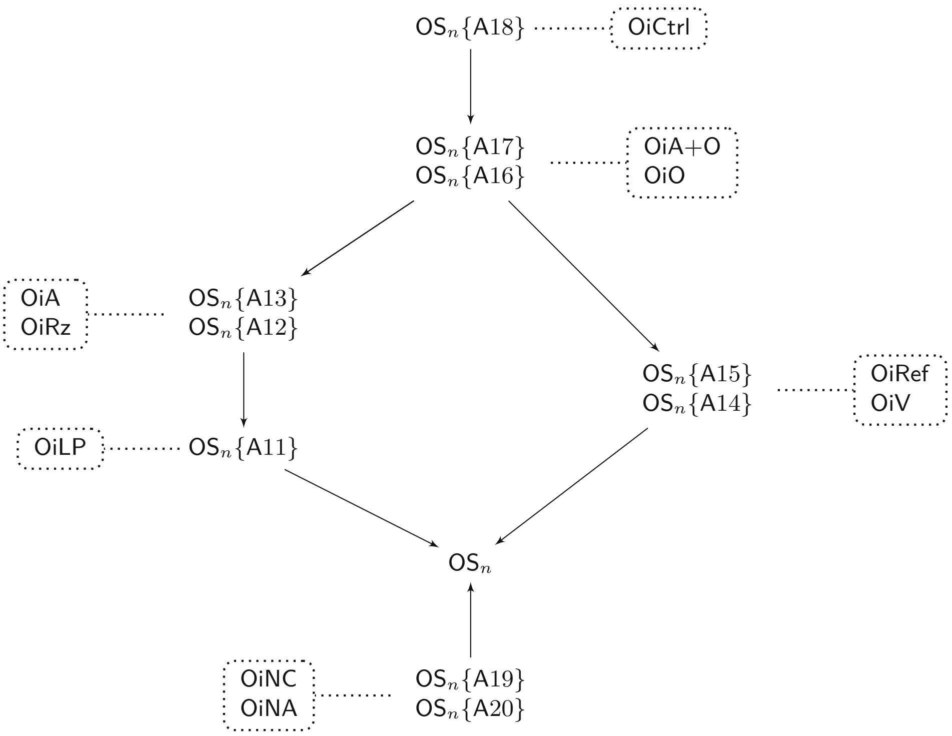

$\mathsf{OS}_n\{\mathsf{A18}\}$
OiCtrl
$\mathsf{OS}_n\{\mathsf{A17}\}$
$\mathsf{OS}_n\{\mathsf{A16}\}$
OiA+O
OiO
OiA
OiRz
$\mathsf{OS}_n\{\mathsf{A13}\}$
$\mathsf{OS}_n\{\mathsf{A12}\}$
$\mathsf{OS}_n\{\mathsf{A15}\}$
$\mathsf{OS}_n\{\mathsf{A14}\}$
OiRef
OiV
OiLP
$\mathsf{OS}_n\{\mathsf{A11}\}$
OS_n
OiNC
OiNA
$\mathsf{OS}_n\{\mathsf{A19}\}$
$\mathsf{OS}_n\{\mathsf{A20}\}$

Figure 3.4: The lattice of $\mathsf{OS}_n\mathsf{X}$ logics for OiC strengthened with the quasi-agentive obligation $\otimes_i\varphi \to \otimes_i[i]\varphi$ (A29). Directed edges point from stronger logics to weaker logics with respect to their expressivity. Reflexive and transitive edges are left implicit. Dotted nodes make explicit which logics correspond to which OiC principle.

logics—referred to as $\mathsf{OS}_n\mathsf{X}$—axiomatizing the ten principles (Objective 1). The logics were subsequently employed to provide a formal taxonomy of OiC, logically determining the (in)dependencies between the various OiC principles (Objective 2). This gave rise to an *endorsement principle* expressing which endorsement of OiC logically commits one to endorse other readings of OiC. We then extended the class of $\mathsf{OS}_n\mathsf{X}$ logics with other metaethical principles—i.e., No Vacuous Commands, Deontic Contingency, Deontic Consistency, and No Deontic Dilemmas—determining their relation to OiC (Objective 3). Whereas No Deontic Dilemmas is logically independent of OiC, we saw that by adopting Deontic Contingency, particular readings of OiC become equivalent, leading to strictly fewer principles. Last, we extended $\mathsf{OS}_n\mathsf{X}$ with restricted forms of monotonicity and aggregation to restore some of the inferential power lost by adopting a non-normal modal approach (Objective 4).

Part II

Action and Normative Reasoning

CHAPTER 4

Norms and Instruments

Since the introduction of deontic logic in the 1950s by von Wright (1951), developments in deontic logic have been guided by the conviction that *action* is a pivotal component of normative reasoning (Castañeda, 1972; von Wright, 1968). In relation to this, a significant development took place in the 1970s: the introduction of Propositional Dynamic Logic (PDL) (Fischer and Ladner, 1979). Modal logics of PDL focus on analyzing complex actions (or programs) and their relation to results. The framework has been adapted to deontic reasoning (Meyer, 1988) and continues to receive attention to the present day (Giordani and Canavotto, 2016; Giordani and Pascucci, 2022; Hughes et al., 2007). The emphasis on action in normative reasoning led to the distinction between two categories of obligation: *ought to be* and *ought to do* (d'Altan et al., 1996; Castañeda, 1972). Obligations of the first category address *states of affairs*, without referring to how the agent obtains such states of affairs. The second category prescribes *actions* to agents without specifying the possible outcomes that the action might produce. We use *norms to be* and *norms to do* as generalizations of the two categories (also including prohibitions).

There is a third category of norms merging both approaches. The category contains norms that describe a normative relation between an action and a goal, where the action serves as an *instrument* for achieving the goal. We propose the name *norms of instrumentality* to characterize obligations and prohibitions of this type. To the best of our knowledge this category has not yet been investigated. Consider the following example:

> It is prohibited to use nonpublic information as an instrument to acquire financial profit on the stock market.

The above prohibition belongs to this third category. It is a simplified representation of the law on 'insider trading'. This prohibition is neither an instance of norms to be nor of norms to do. That is, it is neither prohibited to use nonpublic information nor is it prohibited to acquire financial profit on the stock market. Only as a means to attain financial profit the use of such information is forbidden. Prohibitions of the form expressed above articulate which *actions* may not be employed as *instruments* for achieving particular *goals*. Despite the ubiquity of normative constraints on instrumentality in legal, social, and ethical systems—think of protocols, rules of games, and fairness constraints—an investigation of their philosophical ramifications in formal logic is absent. This work sets out to provide the formal foundations for the analysis of norms of instrumentality.

Objective 1. *Develop a formal logic of actions to represent and analyze norms of instrumentality.*

The dichotomy between norms to be and norms to do is a central theme of Deontic Logic (Hilpinen and McNamara, 2013; Horty, 2001) and forms a key challenge for Normative Multi-agent Systems (NorMAS) (Pigozzi and van der Torre, 2018). One of the main questions is whether the latter is reducible to the former. An immediate question for our endeavor is whether—and if so, to what extent—norms of instrumentality can be reduced to the two aforementioned norm categories.

Objective 2. *Provide a formal comparison of norms to be and norms to do in relation to norms of instrumentality.*

D'Altan et al. (1996) provide an extensive formal analysis of norms to be and norms to do. The formalism employed there brings together Anderson's (1958) reduction of norms of the first category and Meyer's (1988) reduction of norms of the second category. The resulting system is the multi-modal logic referred to as PDeL, i.e., deontic PDL. Anderson's reduction reduces deontic operators to alethic formulae containing *violation constants*, e.g., "a result φ is obligatory when $\neg\varphi$ strictly implies a violation". Meyer's reduction reduces deontic operators to formulae using action modalities and violation constants, e.g., "an action Δ is obligatory when not performing Δ strictly implies a violation".

We introduced a third reduction: the reduction of action modalities in the style of PDL to alethic formulae containing *action constants*, e.g., "action Δ is performed by agent i when the next moment *witnesses* the successful performance of Δ by agent i". The witness is interpreted as a distinctive state of affairs that preserves the idea of actions as first-class citizens in the formal language. The resulting logic facilitates reasoning about agent-dependent actions within the object language and can be used to formally captures different notions of instrumentality. The reduction was first published in (van Berkel and Pascucci, 2018).

This chapter introduces a logic that brings together the above three reductions: The *Logic of Action and Norms*, (LAN, for short). We use this logic to address the above two objectives.

The philosophical foundation of LAN is Georg Henrik von Wright's theory of agency. Von Wright is well-known for his contributions to modal logic and the philosophy of action

(Stoutland, 2010). He is often referred to as one of the founders of the fields of deontic logic and action logic, and his agency theory (1963; 1968) has proven to be a fruitful base for developing action logics (Åqvist, 2002; Segerberg, 1992). What is more, von Wright (1972b) provides and analysis of instrumentality relations in the context of agents.

Contributions. In this chapter, we address the above objectives. Our main contributions are the following:

First, we develop the logic LAN, which brings together the *three reductions* (Anderson, 1958; Meyer, 1988; van Berkel and Pascucci, 2018). The resulting logic extends previous approaches by permitting us to reason with *agent-dependent* actions, as well as *agent-dependent* obligations and prohibitions, within a multi-agent setting. Furthermore, the logic LAN is sound, strongly complete, and decidable.

Second, we propose and investigate various formalizations of norms of instrumentality. In particular, we formally investigate the three norm categories: we pose desiderata describing the relations between them and evaluate their validity vis-à-vis several deontic principles from the literature (cf. metaethical principles in Chapter 3). We illustrate norms of instrumentality and their relation to the other norm categories through the analysis of a formal example.

Third, we discuss how more refined notions of instrumentality, based on von Wright's philosophy of agency, are formalized in an extension of LAN.

Last, the logic LAN is an action logic in the spirit of PDL. The main difference is that we do not adopt modal operators to express (complex) actions but instead use action constants and a single necessity operator to *define* action modalities. We investigate the relation between LAN and (a fragment of) PDL that uses relativized action negation.

Differences. This chapter is based on three articles: (van Berkel and Pascucci, 2018; van Berkel et al., 2020; van Berkel et al., 2022b). All three articles concern the formal treatment of instrumentality and adopt a similar formalism. The content of Sections 4.2-4.5 were first published in (van Berkel et al., 2020). The (formal) analysis of von Wright's theory of agency and instrumentality in Section 4.1 and Section 4.6 derives from (van Berkel and Pascucci, 2018; van Berkel et al., 2022b). The novel contributions of this chapter are the formal comparison of LAN with (a fragment of) PDL and a more extensive discussion of related work.

Outline. In Section 4.1, we provide an overview of von Wright's theory of agency. The logic LAN is defined in Section 4.2 and soundness and completeness are proven in Section 4.3. After that, we employ the logic for a formal analysis of the three norm categories (Section 4.4). In Section 4.5, we formalize an example protocol containing the three norm types. We discuss how to extend LAN to accommodate several notions of instrumentality discussed by von Wright (Section 4.6). Last, we investigate the relation between LAN and PDL in Section 4.7.

An Example: A Hospital's Health and Safety Protocol

In order to clarify the distinct nature of the three types of norms, we provide an example protocol that serves as a benchmark in evaluating our formal framework. Consider the following (artificial) scenario: The Health and Safety Committee of a public hospital recently established a new set of guidelines to govern and redirect the behavior of surgeons and nurses in treating its patients. In particular, motivated by the increased awareness of the dangers of accidental self-inflicted wounds—caused by using sharp tools during surgery—the committee proposed a new policy: the use of scalpels in the operation room is limited to surgeons and prohibited for assisting nurses. The protocol is summed up accordingly:

N1 Surgeons are obliged to use a prescribed scalpel for bringing about necessary incisions during surgery.

N2 Assisting nurses are not allowed to use scalpels during surgery when the situation is not dire.

N3 Nurses and surgeons alike have the obligations to (i) promote the health of their patients, and (ii) preserve hygiene and safety in the operation room.

First, we observe that N1 expresses a norm belonging to the third, novel category of *norms of instrumentality.* Namely, it is an obligation that specifically relates an action as an instrument to a particular outcome: if an incision is required, then the surgeon is obliged to use the scalpel as a means to bring about the incision. N2 is a prohibition subsumed under *norms to do* and holds independently of the instrument's intended purpose. N3 is an obligation of *norms to be* and holds independently of the instruments used to obtain (i) and (ii).

To emphasize the irreducibility of norms of instrumentality to norms to be and norms to do, consider the following: although a surgeon might be obliged to use a scalpel to ensure a required incision, it does not follow that she must use the scalpel independently of its intended purpose (some outcomes obtained by using the scalpels could be prohibited), nor does it mean that she must bring about the incision by any means necessary (some means could be forbidden). In fact, N1 states that the surgeon has *only* the obligation to ensure the required incision *by means of* using the scalpel.

Remark 4.1. *Instrumentality is a general notion that refers to actions serving goals; cf. (Bratman, 1981; Rao and Georgeff, 1995; von Wright, 1972b). Thus, although in the above example reference is made to a tool—i.e., a scalpel—the instrument under consideration is the action of "using the scalpel" (for the purpose of incision).*

The committee makes two additional (naive) *assumptions* in drafting the protocol:

T1 The protocol offers rules of conduct that suffice to resolve all normative conflicts that may arise during surgery.

T2 The protocol assumes that the actions prescribed to the agents can be consistently and jointly performed.

The committee knows that sub-ideal situations can occur, e.g., whenever an employee (in)voluntarily violates an initial rule. To accommodate such scenarios, the committee additionally provides the following contrary-to-duty (CTD) obligation (see Chapter 1 for an introduction):

E1 In case of failing to preserve hygiene standards during surgery (e.g., in the case of self-inflicted wounds), the employee in question is obliged to leave the operation room and call the safety-emergency number immediately.

The purpose of the above rule is to minimize damage in sub-ideal scenarios. Namely, principle E1 prescribes measures to be taken in case of failure to comply with the prescription in N3. We formalize the example in Section 4.5. There, we analyze the consistency of the protocol and apply it to two different scenarios.

4.1 Action and Instrumentality

The philosophical foundation of our envisioned logic is von Wright's general theory of action, as laid out in (von Wright, 1963a; von Wright, 1968; von Wright, 1972b). In this section, we briefly discuss the basic concepts of this theory and refer to the works of Åqvist (2002) and Stoutland (2010) for a more extensive discussion. At the end of this section, we provide some preliminaries concerning von Wright's (1972b) ideas on instrumentality statements. In brief, instrumentality statements express a relationship between a goal and an action as an instrument for obtaining that goal, cf. (Bratman, 1981; Rao and Georgeff, 1995; von Wright, 1972a).

4.1.1 Von Wright's General Theory of Action

According to von Wright (1963a), acting is interfering with the course of nature. Such interference manifests in bringing something about or in preventing something from happening. What is brought about or prevented is a particular *state of affairs*, i.e., a partial description of the world such as "the window is open". To bring about a result p, means to act "in such a manner that the state of affairs that p is the result of one's action" (1963a, p. 13). Likewise, prevention of p indicates that one's action has succeeded in ensuring $\neg p$.

The above concept of action is founded on the notion of *change*. In fact, for von Wright, any theory of agency and action must presuppose an account of change (Åqvist, 2002). He takes change to define a *transition* from an initial state (i.e., the present moment) to an end-state (i.e., a future moment). Such transitions can be either agent-independent

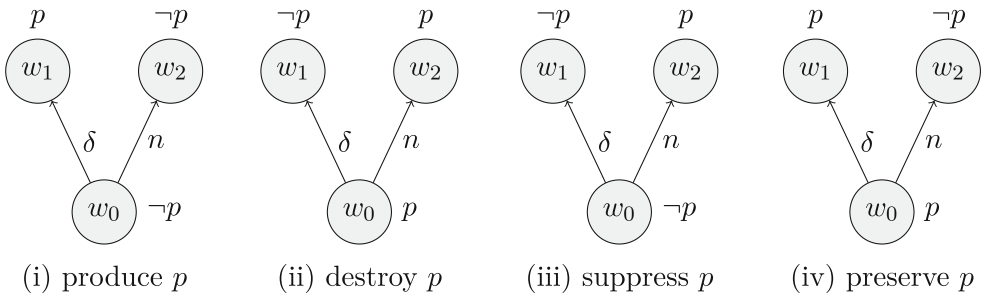


Figure 4.1: Von Wright's four types of elementary action. The node w_0 denotes the initial state and w_1 and w_2 the possible end-states. The transition from w_0 to w_1, denoted by an arrow '→', expresses the performance of the agent's action δ. The alternative transition from w_0 to w_2 indicates the agent's non-interference with nature and is labeled n. The variable p is the state of affairs under consideration.

(e.g., a moon eclipse) or agent-dependent (e.g., me opening the window). The agent-dependent account forces a non-deterministic worldview. That is, to bring something about presupposes at least the following three states: the initial state (in which the agent finds herself), the actual end-state (which is the state that emerges after the performed action), and an alternative end-state (which would result if the agent would refrain from performing the action in question).

Von Wright discusses various relations that may hold between these three states. By bringing together the above account of change with the twofold distinction between bringing about and prevention, he characterizes four types of action: *producing*, *destroying*, *suppressing*, and *preserving*. The first two bring about something, whereas the latter two prevent something. The four types of elementary action are those characterized in Figure 4.1. For instance, at (iii), the act of suppressing p indicates that at the initial state $\neg p$ holds, through the agent's action δ $\neg p$ continues to hold, and if the agent had acted differently, p would have come about. Von Wright's reading of the four action types is arguably too strong since it overlooks the uncertainty of action: for von Wright, the agent's action δ in Figure 4.1 ultimately decides the faith of p. For instance, in the case of producing, by performing δ the agent ensures p, whereas by not-acting the agent can ensure $\neg p$. In other words, von Wright's account takes agency as causally sufficient in both directions (Åqvist, 2002).

Since a general analysis of agency involves many distinct agents and distinct agents can simultaneously perform distinct actions, an individual agent's action is often not causally sufficient. This is known as the uncertainty of action (Horty, 2001) (cf. Section 2.1). We adopt a generalization of von Wright's approach. A transition involves the following three elements: (i) an initial state, (ii) a *set* of actions, and (iii) a *set* of possible final states. Henceforth, we also refer to such states as *moments* in indeterministic time. A single agent does not completely control the course of events, but even the complete set of actions performed by all agents involved does not necessarily entail a unique end-state (cf.

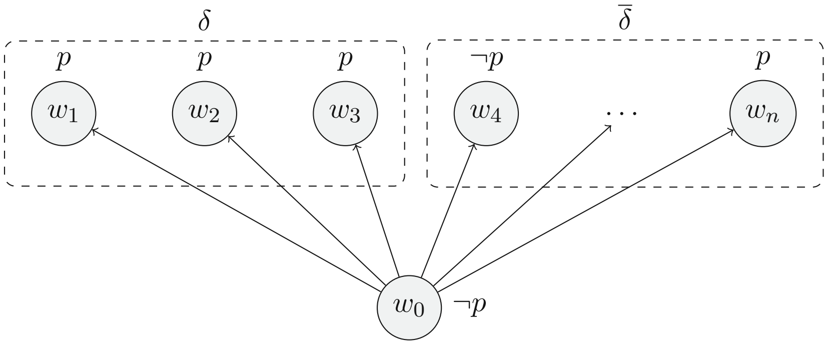


Figure 4.2: Indeterministic time and action. The initial state is w_0, the possible end-states through performing action δ are w_1, w_2, w_3, and those possible through not performing δ (denoted by $\overline{\delta}$) are $w_4, \dots, w_n$. The state of affairs p holds at those moments indicated.

the influence of nature). For this reason, we say that a set of actions *causally contributes to the attainment of* the (actual) end-state of a transition only if the end-state would have been different without the performance of that set of actions.

For instance, in Figure 4.2, we say that the action δ brings the agent from the present moment w_0 to either one of the future moments w_1, w_2, w_3 without strictly determining either of the three. Still, we see that all three moments satisfy p and that performing $\overline{\delta}$—which is the complement of δ—could lead to at least one future moment where $\neg p$ holds, namely, w_4. Thus, we say that the agent can *produce* p by performing δ at w_0, even though the agent cannot secure a unique future moment with action δ.

Last, we follow the usual distinction between *action types* and *action tokens*. The former denote generic categories, such as 'writing', and the latter concern concrete instances in specific circumstances, such as the action of a particular person writing on a particular blackboard on a specific date. See the work of Goldman (1970) for an extensive discussion of types and tokens. Thus, action types can be regarded as categories under which tokens, as individual cases, can be subsumed. von Wright (1963a) adopts a similar demarcation by distinguishing between act-categories, on the one hand, and act-individuals, on the other hand. In this chapter, we consider action types as well as tokens. In the current setting, we restrict ourselves to the analysis of the following types:

- Atomic actions such as "crossing the street";
- Negative actions such as "not crossing the street";
- Complex actions such as disjunctive action, joint action, and sequential action, respectively "turning left or turning right", "turning left and hitting the break", "first turning left and then turning right".

4.1.2 Instrumentality

Central to the study of instrumentality is the relation between an *action* and a *result*. The former is the instrument for the desired outcome expressed in the latter. The outcome can therefore be seen as the purpose of performing the action in question. Thus, we refer to the action as an *instrument* serving a particular *purpose*. An alternative way of referring to this relation is through a *means-end* relation, where the instrument is the means to attain the desired result called end. For example, "pulling down the leaver of a door and drawing the door towards you" is the instrument for the result "the door is open". In Figure 4.2, we find a scenario in which δ is an instrument for purpose p at w_0. Paraphrasing von Wright (1972b, p.21), we say that "δ qualifies as a p-instrument".

We adopt the following definition of a *basic instrumentality relation*:

> *An action Δ is a φ-instrument for agent i at moment w if the performance of Δ by i at w suffices to guarantee the truth of φ.*

In the above definition, we refer to Δ as sufficient for establishing the truth of φ.[1] It differs from necessary means, which are both necessary and sufficient in serving a given purpose. A necessary means is an instrument without which a particular goal cannot be reached. In this chapter, we concentrate on sufficient instruments. Formally, we write

$$[\Delta_i]\varphi$$

to indicate that Δ is a φ-instrument for agent i. Despite its simplicity, the above definition suffices for our formal analysis of norms of instrumentality.

In Section 4.6, we extend our discussion of instrumentality. Although instrumentality statements are critical to practical reasoning (Bratman, 1981; von Wright, 1972a), to the best of our knowledge, von Wright (1972b) is the only philosopher explicitly discussing how such statements can be *obtained, compared, and assessed by agents.* In Section 4.6, we provide a formal discussion of von Wright's analysis.[2]

Remark 4.2. *Means-end reasoning plays a central role in Belief-Desire-Intention (BDI) systems (Rao and Georgeff, 1995). There, means are plans that stipulate a sequence of (sub)actions needed to attain a given goal (where a goal is a consistent, achievable desire) (Rao and Georgeff, 1998). In this chapter, we do not deal with instrumentality in the context of planning (such as in BDI logics). Instead, we investigate instrumentality statements as subject to obligations and prohibitions, i.e., norms of instrumentality. Future research may be directed to investigating the role of such norms in BDI systems.*

[1] This definition of basic instrumentality does not contain a counterfactual element such as found in von Wright's notion of producing. We come back to this in Section 4.2.1.

[2] Furthermore, von Wright is said to be the first to use the term *anankastic conditional*, which refers to means-end conditionals of the form "if you want X, you must do Y". In the *Groundwork for the Metaphysics of Morals* (1785/1999), Kant discusses hypothetical imperatives, which can be seen as anankastic conditionals with imperatives in their consequent. Condoravdi and Lauer (2016) provide a linguistic analysis of the semantics of anankastic conditionals. See (van Berkel et al., 2021b) for a formal analysis of anankastic conditionals in a modal logic setting. We do not pursue this topic in this thesis.

4.2 A Logic of Action and Norms

The analysis in Section 4.1 provides the theoretic foundation of the normal multimodal logic developed in this section: the *$\underline{L}$ogic of $\underline{A}$ction and $\underline{N}$orms* (Objective 1). We write LAN for short. First, we list the fundamental concepts we intend to capture.

Purposes. These are the desired results of actions. We use formulae of arbitrary complexity for their encoding and refer to them using $\varphi, \psi, \chi, \ldots$ (occasionally indexed). We express purposes through descriptions of states of affairs, e.g., "the window is open". Furthermore, descriptions of states of affairs may refer to actions. In those cases, the description functions as a *witness* to the completed performance of an action, e.g., "the door has been opened".

Actions and agents. These are potential instruments for achieving a purpose. We use $\delta, \gamma, \ldots$ (possibly indexed) to represent atomic action types. We build complex action types $\Delta, \Gamma, \ldots$ (possibly indexed) from atomic action types with the use of the following action operations: *action negation* '$-$', *disjunctive action* '$\cup$', and *joint action* '&'. Examples of these operations are given on page 129. Agents perform actions. We denote agents by numbers $i \in \mathbb{N}$. Different actions may be available to different agents. The performance of an *atomic* action-type δ by agent i gives us an action-token $\mathtt{d}_i^{\delta}$ which is a *propositional constant* witnessing the successful performance of the action by that agent. We define a correspondence between action types and action tokens (i.e., witnesses) on page 133.

Reference to possible moments in the immediate future. Action causes change, and change is a temporal transition between moments. In particular, an instrumentality statement refers to how a specific action, as an instrument, may *lead to* a particular state of affairs as its outcome. This 'leading to' is a temporal component referring to possible future moments. We adopt the modal operator $\boxed{\scriptstyle S}$ expressing "in all possible immediate successor moments" (some proposition holds). For instance, let $\mathtt{d}_i^{\delta}$ stand for "the door has been opened by agent i", then the formulae $\boxed{\scriptstyle S}\mathtt{d}_i^{\delta}$ is interpreted as "in all possible immediate successor moments the door has been opened by agent i". We adopt the dual operator $\langle S \rangle$ to denote "in some immediate successor moment" (some proposition holds).

Reference to an immediate actual future. To differentiate between possible successor moments and the actual successor of the moment of evaluation, we adopt the modal operator $\boxed{\scriptstyle A}$, which reads "in the actual immediate successor moment" (some proposition holds). This modality is used to distinguish various concepts of agency, such as cases in which an agent *could* act from those in which she (actually) *will* act.

Norm violations. Following Anderson and Moore (1957), one can reduce obligations and prohibitions to statements about actions provoking sanctions. For instance, "φ is obligatory" can be reduced to "the occurrence of $\neg\varphi$ necessarily implies a sanction". Castañeda (1972) argues that certain problems with this approach are avoided by replacing sanctions with violations. We follow the latter approach. We adopt agent-dependent violations and denote them by propositional constants $\mathtt{v}_i$ for each agent i.

Based on the above list, we define two languages: an action language $\mathcal{L}^{\mathsf{Act}}$, which is an algebra of actions for agent-dependent action types, and the logical language $\mathcal{L}^{\mathsf{LAN}}$ into which these actions are translated. This approach enables us to reason about complex actions from $\mathcal{L}^{\mathsf{Act}}$ as Boolean formulae in the logical language. We use this language to define the central concepts of instrumentality, obligations, and prohibitions in Section 4.4.

Definition 4.1 (Algebra of Actions $\mathcal{L}^{\mathsf{Act}}$)**.** *Let* $\mathsf{Act} = \{\delta, \gamma, \dots\}$ *be a non-empty, countable set of* atomic action-types *and let* $\mathsf{Agents} = \{1, \dots, n\}$ *be a non-empty, finite set of agent labels. The multi-agent language* $\mathcal{L}^{\mathsf{Act}}$ *of complex action types is given via the following BNF grammar:*

$$\Delta ::= \delta_i \mid \Delta \cup \Delta \mid \overline{\Delta}$$

with $\delta \in \mathsf{Act}$ *and* $i \in \mathsf{Agents}$*.*

The above defines combinations of action types as assigned to various agents. We define *joint action* & in terms of action negation and disjunctive action, i.e., $\Delta \& \Gamma := \overline{\overline{\Delta} \cup \overline{\Gamma}}$.[3]

As motivated in the introduction, we employ a reductionist approach to norms via violation constants (Anderson, 1958; Meyer, 1988) and a reduction of actions via action constants (van Berkel and Pascucci, 2018). Let $\mathtt{v}_i$ be a propositional constant witnessing a *norm violation* for agent $i \in \mathsf{Agents}$ and let $\mathsf{Vio} = \{\mathtt{v}_i \mid i \in \mathsf{Agents}\}$ be the set of all agent violation constants. We read $\mathtt{v}_i$ as “agent i has violated a norm”. Furthermore, for any $i \in \mathsf{Agents}$ let $\mathsf{Wit}_i = \{\mathtt{d}_i^\delta, \mathtt{d}_i^\gamma, \dots\}$ be the set of propositional constants that witness the performance of atomic action-types $\delta, \gamma, \dots$ by $i \in \mathsf{Agents}$. We take $\mathtt{d}_i^\delta$ to read “agent i has performed action δ”. We use Wit to denote the set $\bigcup_{i \in \mathsf{Agents}} \mathsf{Wit}_i$. In Definition 4.3, we make the correspondence between agent-dependent action types and propositional constants formally precise. First, we define the language $\mathcal{L}^{\mathsf{LAN}}$.

Definition 4.2 (The Language $\mathcal{L}^{\mathsf{LAN}}$)**.** *Let* $\mathsf{Atoms} = \{p, q, r, \dots\}$ *be a countable set of atomic propositions and let* $\mathsf{Agents} = \{1, \dots, n\}$ *be a non-empty, finite set of agent labels. The language* $\mathcal{L}^{\mathsf{LAN}}$ *is given by the following BNF grammar:*

$$\varphi ::= p \mid \mathtt{v}_i \mid \mathtt{d}_i^\delta \mid \neg\varphi \mid \varphi \vee \varphi \mid \boxed{S}\varphi \mid \boxed{A}\varphi$$

where $p \in \mathsf{Atoms}$*,* $i \in \mathsf{Agents}$*,* $\mathtt{v}_i \in \mathsf{Vio}$ *and* $\mathtt{d}_i^\delta \in \mathsf{Wit}$*.*

The connectives $\wedge$, $\rightarrow$, and $\equiv$ are defined in the usual way. Tautology and contradiction are defined as $\top := p \vee \neg p$, respectively $\bot := p \wedge \neg p$. Formulae of the form $\boxed{S}\varphi$ and $\boxed{A}\varphi$ express, respectively, “in all possible immediate successor moments φ holds” and “in the actual immediate successor moment φ holds”. In what follows, we sometimes omit reference to ‘immediate’. We take $\langle S \rangle$ and $\langle A \rangle$ as the duals of $\boxed{S}$ and $\boxed{A}$, respectively.

[3] Following Åqvist (2002) and von Wright (1951), we adopt the above three action operations. We discuss sequential action in Section 4.7. Both Åqvist and von Wright argue that Boolean operations over actions—including negation—are meaningful. Belnap (1991) claims that actions have no negation.

Definition 4.3 (Translation between $\mathcal{L}^{\mathsf{Act}}$ and $\mathcal{L}^{\mathsf{LAN}}_n$)**.** *The translation t encoding action-types from $\mathcal{L}^{\mathsf{Act}}$ into agent-indexed formulae of $\mathcal{L}^{\mathsf{LAN}}$ is established recursively:*

- *For any $\delta_i \in \mathcal{L}^{\mathsf{Act}}$, $t(\delta_i) = \mathsf{d}_i^{\delta}$, with $\mathsf{d}_i^{\delta} \in \mathcal{L}^{\mathsf{LAN}}$;*
- *For any $\Delta \in \mathcal{L}^{\mathsf{Act}}$, $t(\overline{\Delta}) = \neg t(\Delta)$;*
- *For any $\Delta, \Gamma \in \mathcal{L}^{\mathsf{Act}}$, $t(\Delta \cup \Gamma) = t(\Delta) \vee t(\Gamma)$.*

The advantage of this translation is that it enables us to reason with actions in the logical language while simultaneously distinguishing such formulae from other (non-action) formulae in the language. This distinction will prove beneficial for defining various agentive and deontic modalities and axiomatizing action-specific properties.

Remark 4.3 (Individual Agency)**.** *The language $\mathcal{L}^{\mathsf{Act}}$ defines multi-agent expressions that arbitrarily combine actions of various agents, e.g., $(\delta_i \cup \chi_j)\&\overline{\gamma_k}$ for $i, j, k \in \mathsf{Agents}$. In the remainder of this chapter, we are mainly interested in individual agency, i.e.,* complex actions performed by a single agent*. To accommodate this, we adopt the following notation: we write $\Delta_i \in \mathcal{L}^{\mathsf{Act}}$ whenever all atomic action types occurring in Δ are labeled with agent i. We also say that Δ_i is an i-dependent action of type Δ.*

The translation enables us to define the notion of basic φ-instruments (page 130) through a reduction to boolean formulae of action constants and the immediate successor relation:

$$[\Delta_i]\varphi := \boxed{\mathsf{S}}(t(\Delta_i) \rightarrow \varphi)$$

There is a strong connection between formulae of the form $[\Delta_i]\varphi$ and formulae used in languages of Propositional Dynamic Logic (PDL) (Fischer and Ladner, 1979). When used in combination with actions, the modality $\boxed{\mathsf{S}}$ may be taken as an indeterministic execution operator in the spirit of PDL: namely, "every successful execution of Δ, guarantees φ". The approach undertaken in this chapter can be seen as a reduction of PDL-like logics to alethic modal logic with action constants, similar to the Andersonian reduction of translating deontic formulae into alethic modal formulae with violation constants. We discuss the relation between our logic and PDL in more detail in Section 4.7.

4.2.1 Agentive Modalities in LAN

To illustrate the potential of our formal language, we discuss various agentive notions that can be defined in $\mathcal{L}^{\mathsf{LAN}}$. We use these definitions to formalize the example protocol in Section 4.5. The first three modalities define the notions of *would*, *could*, and *will*:

Would

$$[\Delta_i]^{would}\varphi := \boxed{\mathsf{S}}(t(\Delta_i) \rightarrow \varphi) \qquad (\mathsf{d}1)$$

Could

$$[\Delta_i]^{could}\varphi := \boxed{\mathsf{S}}(t(\Delta_i) \to \varphi) \wedge \Diamond_{\mathsf{S}}\, t(\Delta_i) \tag{d2}$$

Will

$$[\Delta_i]^{will}\varphi := \boxed{\mathsf{S}}(t(\Delta_i) \to \varphi) \wedge \Diamond_{\mathsf{A}}\, t(\Delta_i) \tag{d3}$$

The formula $[\Delta_i]^{would}\varphi$ (d1) means that "at the current state, by performing Δ, i would bring about φ".[4] The formula $[\Delta_i]^{could}\varphi$ (d2) means that "at the moment of evaluation, by performing Δ, i would bring about φ and i could (i.e., is able to) perform Δ". Finally, the formula $[\Delta_i]^{will}\varphi$ (d3) means that "at the moment of evaluation, by performing Δ, i would bring about φ and i will perform Δ". Although we can define multi-agent variants such as $[\Delta_i \& \Gamma_j]^{could}\varphi$—referring to the agents i and j's ability to jointly secure φ—we focus on single-agent actions for now (see Remark 4.3).

Furthermore, we can formally define agentive modalities corresponding to von Wright's four elementary action types (Section 4.1). Von Wright's action types are *deliberative* in nature by excluding trivial outcomes (e.g., $\top$) and ensuring that outcomes are about contingent states of affairs φ, namely, for which $\Diamond_{\mathsf{S}}\varphi$ and $\Diamond_{\mathsf{S}}\neg\varphi$ hold.[5] Recall from our discussion on causal contribution on page 129, that we take a slightly weaker standpoint than von Wright's. Following Åqvist (2002), we say that agent i produces p by performing Δ if the action Δ suffices to bring about p and not performing Δ *may* result in $\neg p$. This is reflected in definitions (d4), (d5), (d6), and (d7) below.[6]

Produce

$$[\Delta_i]^{prod}p := \neg p \wedge [\Delta_i]^{will}p \wedge \Diamond_{\mathsf{S}}\neg p \tag{d4}$$

Destroy

$$[\Delta_i]^{destr}p := p \wedge [\Delta_i]^{will}\neg p \wedge \Diamond_{\mathsf{S}} p \tag{d5}$$

Suppress

$$[\Delta_i]^{supp}p := \neg p \wedge [\Delta_i]^{will}\neg p \wedge \Diamond_{\mathsf{S}} p \tag{d6}$$

Preserve

$$[\Delta_i]^{pres}p := p \wedge [\Delta_i]^{will}p \wedge \Diamond_{\mathsf{S}}\neg p \tag{d7}$$

[4] The notion of 'would' equals the basic instrumentality relation. Namely, action Δ is a φ-instrument for agent i at a moment, whenever every performance of Δ by i at this moment would bring about φ. This suffices for now. In Section 4.6, we present some more involved definitions of instrumentality.

[5] See Section 3.5 for a discussion of trivial outcomes and contingency in deontic STIT logic.

[6] We define the four elementary action types relative to propositional atoms. The use of arbitrary formulae φ would make (d4) and (d5) equivalent. The same applies to (d6) and (d7). Consequently, a generalization of these four action types to arbitrary formulae would result in von Wright's two general categories of "bringing about something", respectively "preventing something" (see page 128).

By using the notions of 'would' (d1) and 'could' (d2), one can define variations of the above four action types. For instance, (d8) expresses the idea that "agent i could destroy p by performing the action Δ". The first conjunct of (d8) states that p is presently the case, the second ensures that by performing Δ agent i would bring about $\neg p$ and i could perform Δ, and last, it is possible that p is not destroyed.

Could Destroy

$$[\Delta_i]^{could}_{destr} p := p \wedge [\Delta_i]^{could} \neg p \wedge \Diamond_{\mathsf{S}} p \tag{d8}$$

Last, consider *forbearance*. This agentive notion expresses more than merely not acting. Forbearing assumes the agent's *ability* to perform the forborne action. In (d9), the occurrence of $\top$ denotes that forbearance refers to action irrespective of its outcome.

Forbear

$$[\Delta_i]^{forb} \top := [\Delta_i]^{could} \top \wedge [\overline{\Delta}_i]^{will} \top \tag{d9}$$

Definition (d9) reads "agent i forbears performing action Δ whenever i could perform action Δ, but will instead perform the action's complement $\overline{\Delta}$". One can see how the notion of forbearance can be extended to incorporate the four elementary action types. In Section 4.4, we show how this language can be used to express various deontic notions.

4.2.2 Axiomatization of LAN

The Hilbert-style axiomatization of the logic LAN is given below.

Definition 4.4 (The Axiomatization of LAN)**.** *The logic* LAN *is axiomatized by the following collection of axiom schemes and rules:*

A0. *All classical propositional tautologies;*

R0. *From* φ *and* $\varphi \rightarrow \psi$*, infer* ψ*;*

A1. $\boxed{\mathsf{S}}(\varphi \rightarrow \psi) \rightarrow (\boxed{\mathsf{S}}\varphi \rightarrow \boxed{\mathsf{S}}\psi)$*;*

A2. $\boxed{\mathsf{A}}(\varphi \rightarrow \psi) \rightarrow (\boxed{\mathsf{A}}\varphi \rightarrow \boxed{\mathsf{A}}\psi)$*;*

A3. $\Diamond_{\mathsf{A}}\varphi \rightarrow \boxed{\mathsf{A}}\varphi$*;*

A4. $\boxed{\mathsf{S}}\varphi \rightarrow \boxed{\mathsf{A}}\varphi$*;*

A5. *For any* $1, \dots, n \in \mathsf{Agents}$ *and* $\Delta_1, \dots, \Gamma_n \in \mathcal{L}^{\mathsf{Act}}$, $(\Diamond_{\mathsf{S}} t(\Delta_1) \wedge \dots \wedge \Diamond_{\mathsf{S}} t(\Gamma_n)) \rightarrow \Diamond_{\mathsf{S}}(t(\Delta_1) \wedge \dots \wedge t(\Gamma_n))$*;*

A6. $\Diamond_{\mathsf{S}} \mathtt{v}_i \rightarrow \Diamond_{\mathsf{S}} \neg \mathtt{v}_i$*;*

R1. *From* φ*, infer* $\boxed{\scriptstyle S}\varphi$*;*

where we have a copy of A6 *for each* $i \in \mathsf{Agents}$*. The* logic LAN *is the smallest set of formulae from* $\mathcal{L}^{\mathsf{LAN}}$ *closed under all instances of the axiom schemes and applications of the inference rules* R0 – R1*. Whenever* $\varphi \in \mathsf{LAN}$*, we say that* $\varphi \in \mathcal{L}^{\mathsf{LAN}}$ *is a* LAN-theorem *and write* $\vdash_{\mathsf{LAN}} \varphi$*. Last,* LAN*-derivabiliry is defined as usual (see Definition 2.3).*

The axioms A1, A2, A4, and R1 specify that both $\boxed{\scriptstyle S}$ and $\boxed{\scriptstyle A}$ are normal modal operators. In addition, axiom A3 ensures that every moment has at most one actual successor. Axiom A4 guarantees that every actual successor moment is also a possible successor moment.[7] Axiom A5 captures a principle called 'independence of agents'. It ensures that if an agent can perform a particular action at a specific moment in time, that agent can perform that action irrespective of the actions performed by any of the other agents. In other words, any combination of actions available to different agents at a given moment is consistent. Independence of agents is a fundamental property of the agency formalism called STIT logic (Belnap et al., 2001) (see Chapter 2). We adopt this property to the setting where we have explicit actions available to agents. In the context of LAN, independence of agents additionally ensures that, if an action Δ is a φ-instrument for i, then failing to produce φ by performing Δ cannot be caused by the interference of other agents. In other words, Δ is a proper φ instrument for agent i. We point out that axiom A5 requires the agents $1, \dots, n$ to be distinct, whereas the agent-independent actions $\Delta, \dots, \Gamma$ as performed by $1, \dots, n$, respectively, may or may not be distinct. Last, A6 enforces that if there is a possible future in which a norm violation occurs for some agent, then there is also an alternative future available in which a norm violation is avoided for that agent. This last condition is in the spirit of the *deontic contingency principle* proposed by Anderson and Moore (1957). There, a principle is adopted ensuring that it is both possible to violate a norm and possible to avoid such a violation. Observe that A6 is phrased as a conditional property, which is weaker than the one by Anderson and Moore (1957) (see Section 3.6 for a discussion of deontic contingency).

4.2.3 Semantics for LAN

We adopt relational semantics to characterize LAN. First, we define some preliminaries.

[7] Branching time structures are a standard solution to modeling indeterministic scenarios (Belnap et al., 2001; Thomason, 1984). In such representations of time, finding an appropriate way of referring to *actuality* is not straightforward. In indeterministic time, every moment may have several possible future continuations. The actions performed by the agents reduce the possible futures. Conceptually, the actual successor modality allows us to single out that successor moment that actually obtains, e.g., as the result of the actions performed by the involved agents. We note that the approach adopted here does not solve the philosophical issues of actuality in branching time structures. However, we point out that the use of the actual successor modality $\boxed{\scriptstyle A}$ is not required for the analysis of instrumentality in Section 4.4 and is primarily used to model the agentive concept of "will" in Section 4.2 and the example scenario in Section 4.5. We refer to (Belnap et al., 2001; Belnap and Green, 1994) for a discussion of the problems associated with actuality.

Definition 4.5. *Let W be a non-empty set of worlds $w, v, u, \dots$ and let for each $\mathtt{d}_i^\delta \in \mathsf{Wit}$, $W_{\mathtt{d}_i^\delta} \subseteq W$ be a subset of worlds. For arbitrary $\Delta, \Gamma \in \mathcal{L}^{\mathsf{Act}}$ we define the set $W_{t(\Delta)}$ using the following recursive clauses:*

- $W_{t(\delta_i)} := W_{\mathtt{d}_i^\delta}$;
- $W_{t(\overline{\Delta})} := W \setminus W_{t(\Delta)}$;
- $W_{t(\Delta \sqcup \Gamma)} := W_{t(\Delta)} \cup W_{t(\Gamma)}$.

We write $W_{t(\Delta_i)}$ whenever $\Delta \in \mathcal{L}^{\mathsf{Act}}$ is an agent i-dependent action *(see Remark 4.3).*

The above defines sets of worlds that witness action types. We use this definition for capturing independence of agents for agent-dependent action types. In Lemma 4.1 we show that this recursive definition is well-defined.

Definition 4.6 (Frames and Models for LAN)**.** *A* LAN-frame *is defined as a tuple $\mathfrak{F} = \langle W, \{W_{\mathtt{d}_i^\delta} \mid \mathtt{d}_i^\delta \in \mathcal{L}^{\mathsf{LAN}}\}, \{W_{\mathtt{v}_i} \mid \mathtt{v}_i \in \mathcal{L}^{\mathsf{LAN}}\}, \mathcal{R}_{\boxed{\mathsf{S}}}, \mathcal{R}_{\boxed{\mathsf{A}}}\rangle$. Let $\mathcal{R}_{[\alpha]} \subseteq W \times W$ and $\mathcal{R}_{[\alpha]}(w) := \{v \in W \mid (w, v) \in \mathcal{R}_{[\alpha]}\}$ for $[\alpha] \in \{\boxed{\mathsf{S}}, \boxed{\mathsf{A}}\}$. Let W be a non-empty set of worlds $w, v, u, \dots$ such that the following hold:*

R1 *For each $\mathtt{d}_i^\delta \in \mathsf{Wit}$, $W_{\mathtt{d}_i^\delta} \subseteq W$;*

R2 *For each $\mathtt{v}_i \in \mathsf{Vio}$, $W_{\mathtt{v}_i} \subseteq W$;*

R3 *For all $w, u, v \in W$, if $u \in \mathcal{R}_{\boxed{\mathsf{A}}}(w)$ and $v \in \mathcal{R}_{\boxed{\mathsf{A}}}(w)$, then $u = v$;*

R4 *For all $w, v \in W$, if $v \in \mathcal{R}_{\boxed{\mathsf{A}}}(w)$, then $v \in \mathcal{R}_{\boxed{\mathsf{S}}}(w)$;*

R5 *For all $w, \in W$, all $1, \dots, n \in \mathsf{Agents}$ with $\Delta_1, \dots, \Gamma_n \in \mathcal{L}^{\mathsf{Act}}$, if $u_1, \dots, u_n \in \mathcal{R}_{\boxed{\mathsf{S}}}(w)$ and $u_1 \in W_{t(\Delta_1)}, \dots, u_n \in W_{t(\Gamma_n)}$, then there is a world $v \in W$ such that $v \in \mathcal{R}_{\boxed{\mathsf{S}}}(w)$ and $v \in W_{t(\Delta_1)} \cap \dots \cap W_{t(\Gamma_n)}$;*

R6 *For all $w \in W$ and all $i \in \mathsf{Agents}$, if there exists a $v \in W$ such that $v \in \mathcal{R}_{\boxed{\mathsf{S}}}(w)$ and $v \in W_{\mathtt{v}_i}$, then there is a world $u \in W$ such that $u \in \mathcal{R}_{\boxed{\mathsf{S}}}(w)$ and $u \in W \setminus W_{\mathtt{v}_i}$;*

A LAN-model *is a tuple $\mathfrak{M} = \langle \mathfrak{F}, V\rangle$ where $\mathfrak{F}$ is a LAN-frame and V is a valuation function mapping propositional atoms and constants to subsets of W, i.e., $V : \mathsf{Atoms} \cup \mathsf{Wit} \cup \mathsf{Vio} \mapsto \mathcal{P}(W)$. The following two restrictions hold:*

- $V(\mathtt{d}_i^\delta) = W_{\mathtt{d}_i^\delta}$*, for any* $\mathtt{d}_i^\delta \in \mathsf{Wit}$*;*
- $V(\mathtt{v}_i) = W_{\mathtt{v}_i}$*, for any* $\mathtt{v}_i \in \mathsf{Vio}$*.*

In what follows, we use the terms 'worlds' and 'moments' interchangeably. The conditions **R1** and **R2** stipulate that the sets $W_{\mathtt{d}_i^\delta}$, respectively $W_{\mathtt{v}_i}$ contain moments from W that witness the performance of the atomic action $\mathtt{d}_i^\delta$, respectively the violation of a norm for agent i. The restrictions on the valuation function V ensure that those moments witnessing $\mathtt{d}_i^\delta$ and $\mathtt{v}_i$ *satisfy* those constants (cf. Definition 4.7 below). In other words, in LAN-models, the valuation of constants is fixed on the level of LAN-frames. Consequently, the semantic interpretation of constants is fixed for every model defined over such a frame. (That this observation generalizes to arbitrary actions is demonstrated in Lemma 4.1.) This particular feature enables us to provide frame properties (and corresponding axioms) that characterize the behavior of actions and violations (cf. hybrid logics (Braüner, 2022)). The binary relation $\mathcal{R}_{\boxed{\mathsf{S}}}$ represents *possible* immediate transitions from the current moment, and the relation $\mathcal{R}_{\boxed{\mathsf{A}}}$ represents the *actual* transition from the current moment. Property **R3** stipulates that for each moment, there is at most one actual future moment (cf. $\mathsf{A3}$). **R4** ensures that the actual future moment must be one of the possible future moments (cf. $\mathsf{A4}$). Furthermore, **R5** expresses independence of agents (cf. $\mathsf{A5}$). Last, condition **R6** expresses the *weak contingency* principle for agent-dependent norm violations (cf. $\mathsf{A6}$).

The semantic interpretation of $\mathcal{L}^{\mathsf{LAN}}$ is defined below.

Definition 4.7 (Semantics of LAN-models)**.** *Let $\mathfrak{M}$ be a LAN-model and let $w \in W$ of $\mathfrak{M}$. The* satisfaction *of a formula $\varphi \in \mathcal{L}^{\mathsf{LAN}}$ in $\mathfrak{M}$ at w is defined accordingly:*

1. $\mathfrak{M}, w \models p$ *iff* $w \in V(p)$;
2. $\mathfrak{M}, w \models \mathtt{d}_i^\delta$ *iff* $w \in V(\mathtt{d}_i^\delta) = W_{\mathtt{d}_i^\delta}$;
3. $\mathfrak{M}, w \models \mathtt{v}_i$ *iff* $w \in V(\mathtt{v}_i) = W_{\mathtt{v}_i}$;
4. $\mathfrak{M}, w \models \neg\varphi$ *iff not* $\mathfrak{M}, w \models \varphi$;
5. $\mathfrak{M}, w \models \varphi \vee \psi$ *iff* $M, w \models \varphi$ *or* $\mathfrak{M}, w \models \psi$;
6. $\mathfrak{M}, w \models \boxed{\mathsf{S}}\varphi$ *iff for all* $v \in \mathcal{R}_{\boxed{\mathsf{S}}}(w)$, $\mathfrak{M}, v \models \varphi$;
7. $\mathfrak{M}, w \models \boxed{\mathsf{A}}\varphi$ *iff for all* $v \in \mathcal{R}_{\boxed{\mathsf{A}}}(w)$, $\mathfrak{M}, v \models \varphi$.

Global truth, validity, and semantic entailment are defined as usual (see Definition 2.5).

The following lemma states that the set $W_{t(\Delta_i)}$ of Definition 4.5 is well-defined.

Lemma 4.1. *For each $\Delta \in \mathcal{L}^{\mathsf{Act}}$, each LAN-model $\mathfrak{M}$, and each moment $w \in W$ of $\mathfrak{M}$, we have: $\mathfrak{M}, w \models t(\Delta)$ iff $w \in W_{t(\Delta)}$.*

Proof. The proof is by induction on the complexity of Δ. We omit reference to $\mathfrak{M}$. *Base case.* Let $\Delta = t(\delta_i) = \mathtt{d}_i^{\delta}$. Then, $w \models t(\delta_i)$ iff $w \in W_{\mathtt{d}_i^{\delta}}$ follows directly from the definition of $W_{\mathtt{d}_i^{\delta}}$ and V. *Inductive step.* We consider the two complex action operators:

($\Delta = \overline{\Gamma}$) Left-to-Right. Assume $w \models t(\overline{\Gamma})$. By the translation function, this means $w \models \neg t(\Gamma)$. Hence, we have $w \not\models t(\Gamma)$. By IH this gives us $w \notin W_{t(\Gamma)}$. Hence, $w \in W \setminus W_{t(\Gamma)} = W_{t(\overline{\Gamma})}$. Right-to-Left. Assume $w \in W_{t(\overline{\Gamma})}$. This means $w \in W \setminus W_{t(\Gamma)}$ and so $w \notin W_{t(\Gamma)}$. By IH this means $w \not\models t(\Gamma)$. Which, gives us $w \models \neg t(\Gamma)$ and by the definition of translation function this implies $w \models t(\overline{\Gamma})$.

($\Delta_i = \Gamma^1 \cup \Gamma^2$) Left-to-Right. By the translation function, $w \models t(\Gamma^1 \cup \Gamma^2)$ iff $w \models t(\Gamma^1) \vee t(\Gamma^2)$. By semantic definition, $w \models t(\Gamma^1) \vee t(\Gamma^2)$ iff $w \models t(\Gamma^1)$ or $w \models t(\Gamma^2)$. By IH, $w \models t(\Gamma^1)$ or $w \models t(\Gamma^2)$ iff $w \in W_{t(\Gamma^1)}$ or $w \in W_{t(\Gamma^2)}$. By definition of sets, $w \in W_{t(\Gamma^1)}$ or $w \in W_{t(\Gamma^2)}$ iff $w \in W_{t(\Gamma^1)} \cup W_{t(\Gamma^2)}$. By Definition 4.5, $w \in W_{t(\Gamma^1)} \cup W_{t(\Gamma^2)}$ iff $w \in W_{t(\Gamma^1 \cup \Gamma^2)}$. QED

Corollary 4.1. *In the logic* LAN*, we have for each* $\Delta, \Gamma \in \mathcal{L}^{\mathsf{Act}}$*:* $t(\Delta \& \Gamma) = t(\Delta) \wedge t(\Gamma)$ *and* $W_{t(\Delta\&\Gamma)} = W_{t(\Delta)} \cap W_{t(\Gamma)}$*.*

Furthermore, as a consequence of Lemma 4.1, we obtain the following semantic interpretation of the defined action modality $[\Delta]$:

$$\mathfrak{M}, w \models [\Delta]\varphi \text{ iff for all } v \in \mathcal{R}_{\boxed{S}}(w), \text{ if } v \in W_{t(\Delta)} \text{ then } \mathfrak{M}, v \models \varphi$$

In other words, every immediate successor witnessing the performance of Δ by an agent i guarantees the truth of φ.

Remark 4.4. *The following logical relation holds between the agentive modalities 'would', 'could', and 'will' in the logic* LAN*:* $\models [\Delta_i]^{will}\varphi \to [\Delta_i]^{could}\varphi$ *and* $\models [\Delta_i]^{could}\varphi \to [\Delta_i]^{would}\varphi$*. The proof of the first claim is a consequence of property* ***R4****. The second claim follows directly from the definitions of these modalities.*

Example 4.1. *Consider the single-agent* LAN*-model presented in Figure 4.3. Let* $\mathsf{Agents} = \{i\}$, $\mathsf{Atoms} = \{p\}$, $\mathsf{Act} = \{\delta\}$, $\mathsf{Vio} = \{\mathtt{v}_i\}$, *and* $\mathsf{Wit} = \{\mathtt{d}_i^{\delta}\}$*. We define the model* $\mathfrak{M}$ *as follows:* $W = \{w_0, w_1, w_2, w_3, w_4, \ldots, w_n\}, \mathcal{R}_{\boxed{S}} = \{(w_0, w_i) \mid 1 \leq i \leq n\}$, $\mathcal{R}_{\boxed{A}} = \{(w_0, w_4)\}$, $V(\mathtt{d}_i^{\delta}) = W_{\delta_i} = \{w_1, w_2, w_3\}$, $W_{\mathtt{v}_i} = V(\mathtt{v}_i) = \emptyset$*, and* $V(p) = W \setminus \{w_4\}$*. It can be straightforwardly checked that* $\mathfrak{M}$ *is a* LAN*-model. We list some observations: at* w_0 *it is true that by performing* δ *agent* i could *produce* p*, i.e.,* $\mathfrak{M}, w_0 \models \neg p \wedge [\delta_i]^{could} p \wedge \mathord{\Diamond\mkern-11mu{\scriptstyle S}}\,\neg p$ *(cf. Section 4.2.1). However, since* w_4 *is the actual successor of* w_0*, agent* i *will not produce* p*. For this, it suffices to observe that* $\mathfrak{M}, w_0 \not\models [\delta_i]^{will} p$ *due to* $\mathfrak{M}, w_0 \models \mathord{\Diamond\mkern-11mu{\scriptstyle A}}\,\neg \mathtt{d}_i^{\delta}$*. In Section 4.4, we consider deontic examples that employ violation constants.*

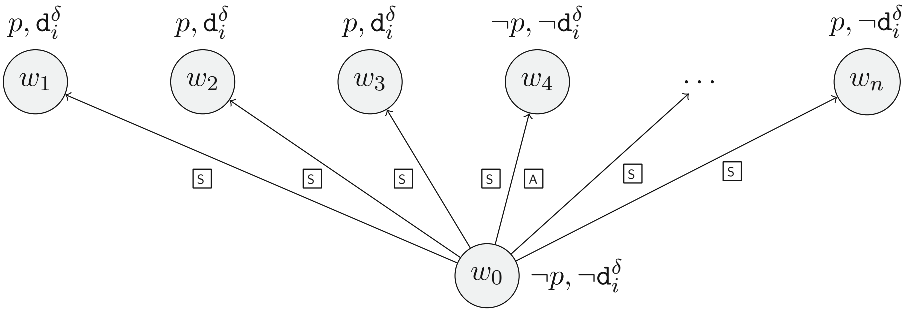


Figure 4.3: The single-agent LAN-model of Example 4.1.

4.3 Soundness and Completeness

Soundness of LAN is demonstrated as usual. Completeness is shown through adopting the canonical model approach adapted to the use of propositional constants. Last, we note that LAN was shown decidable in (van Berkel et al., 2020). We refer to this work for its proof.

Theorem 4.1 (Soundness of LAN)**.** *For any formula $\varphi \in \mathcal{L}^{\mathsf{LAN}}$, and any $\Gamma \subseteq \mathcal{L}^{\mathsf{LAN}}$: if $\Gamma \vdash_{\mathsf{LAN}} \varphi$, then $\Gamma \models_{\mathsf{LAN}} \varphi$.*

Proof. It suffices to demonstrate that all axioms are LAN-valid, and the logical rules of LAN preserve truth on the respective frame classes. Take an arbitrary LAN-model $\mathfrak{M}$ and an arbitrary $w \in W$ of $\mathfrak{M}$. In what follows, we omit reference to $\mathfrak{M}$. The axiom schemes $\mathsf{A0}$, $\mathsf{A1}$, and $\mathsf{A2}$, and rules $\mathsf{R0}$ and $\mathsf{R1}$ are valid, respectively preserve validity on all relational frames (Blackburn et al., 2004). We omit their proofs.

A3 $\langle\mathsf{A}\rangle\varphi \to \boxed{\mathsf{A}}\varphi$. Assume $\mathfrak{M}, w \models \langle\mathsf{A}\rangle\varphi$. By semantic definition of $\langle\mathsf{A}\rangle$, there is a $v \in \mathcal{R}_{\boxed{\mathsf{A}}}(w)$ such that $\mathfrak{M}, v \models \varphi$. Take an arbitrary $u \in \mathcal{R}_{\boxed{\mathsf{A}}}(w)$. By **R3** we know that $v = u$ and so, $\mathfrak{M}, u \models \varphi$. Since u was arbitrary we have $\mathfrak{M}, w \models \boxed{\mathsf{A}}\varphi$.

A4 $\boxed{\mathsf{S}}\varphi \to \boxed{\mathsf{A}}\varphi$. Assume $\mathfrak{M}, w \models \boxed{\mathsf{S}}\varphi$. By semantic definition of $\boxed{\mathsf{S}}$ we know that for all $v \in \mathcal{R}_{\boxed{\mathsf{S}}}(w)$, $\mathfrak{M}, v \models \varphi$. By clause **R4** we know that $\mathcal{R}_{\boxed{\mathsf{A}}}(w) \subseteq \mathcal{R}_{\boxed{\mathsf{S}}}(w)$, and so for all $v \in \mathcal{R}_{\boxed{\mathsf{A}}}(w)$, $\mathfrak{M}, v \models \varphi$ too. Hence, $\mathfrak{M}, w \models \boxed{\mathsf{A}}\varphi$.

A5 For any $1, \dots, n \in \mathsf{Agents}$ and $\Delta_1, \dots, \Delta_n \in \mathcal{L}^{\mathsf{Act}}$, $(\langle\mathsf{S}\rangle t(\Delta_1) \wedge \dots \wedge \langle\mathsf{S}\rangle t(\Gamma_n)) \to \langle\mathsf{S}\rangle(t(\Delta_1) \wedge \dots \wedge t(\Delta_n)$. Assume $1, \dots, n \in \mathsf{Agents}$, $\Delta_1, \dots, \Gamma_n \in \mathcal{L}^{\mathsf{Act}}$, and $\mathfrak{M}, w \models \langle\mathsf{S}\rangle t(\Delta_1) \wedge \dots \wedge \langle\mathsf{S}\rangle t(\Gamma_n)$. Consequently, by Lemma 4.1, we can infer that there are $v_1, \dots, v_n \in \mathcal{R}_{\boxed{\mathsf{S}}}(w)$ such that $v_1 \in W_{t(\Delta_1)}, \dots, v_n \in W_{t(\Gamma_n)}$. By **R5**, we know that there is $u \in \mathcal{R}_{\boxed{\mathsf{S}}}(w)$ and $u \in W_{t(\Delta_1)} \cap \dots \cap W_{t(\Gamma_n)}$. Hence, $\mathfrak{M}, w \models \langle\mathsf{S}\rangle(t(\Delta_1) \wedge \dots \wedge t(\Gamma_n))$.

A6 $\langle\mathsf{S}\rangle\mathtt{v}_i \to \langle\mathsf{S}\rangle\neg\mathtt{v}_i$. Assume $w \models \langle\mathsf{S}\rangle\mathtt{v}_i$ for some $\mathtt{v}_i \in \mathcal{L}^{\mathsf{LAN}}$. Then, there is a $v \in W$ such that $v \in \mathcal{R}_{\boxed{\mathsf{S}}}(w)$ and $w \models \mathtt{v}_i$. By the semantic definition of $\mathtt{v}_i$, this means $v \in W_{\mathtt{v}_i}$.

By **R6**, there is a $u \in W$ such that $u \in \mathcal{R}_{\boxed{\mathsf{S}}}(w)$ and $u \in W \setminus W_{\mathsf{v}_i}$. By semantic definition, this means $w \models \Diamond\!\!\!\!\!{\scriptstyle\mathsf{S}}\;\neg\mathsf{v}_i$.

The above holds for each $i \in \mathsf{Agents}$ and each $\Delta_j \in \mathcal{L}^{\mathsf{Act}}$. Strong soundness follows through reasoning similar to Theorem 2.1 on page 39 (Chapter 2). QED

We adapt the method of canonical models for normal modal logics (Blackburn et al., 2004) for proving completeness. The strategy is as follows: we define the notion of a LAN-maximally consistent set of $\mathcal{L}^{\mathsf{LAN}}$ formulae (Definition 4.8). These sets are used as worlds in constructing a canonical model for the logic LAN (Definition 4.9). Subsequently, we prove a truth lemma (Lemma 4.5), ensuring that every LAN-consistent set of formulae can be satisfied on this canonical model. The main aim is to demonstrate that the obtained canonical model is an LAN-model (Lemma 4.7). Last, this model is used to prove strong completeness via contraposition.

We reserved the notation $\Delta, \Gamma, \ldots$ for arbitrary action types of $\mathcal{L}^{\mathsf{Act}}$. To enhance clarity in the completeness proof, we reserve $\Sigma, \Theta, \Pi, \ldots$ for LAN-CSs and LAN-MCSs.

Definition 4.8 (LAN-CS and LAN-MCS)**.** *A set $\Sigma \subset \mathcal{L}^{\mathsf{LAN}}$ is a* LAN-consistent *(*LAN*-CS) iff $\Sigma \nvdash_{\mathsf{LAN}} \bot$. A set $\Sigma \subset \mathcal{L}^{\mathsf{LAN}}$ is an* LAN-maximally consistent *(*LAN*-MCS) iff Σ is an* LAN*-CS and for any set $\Sigma' \subseteq \mathcal{L}^{\mathsf{LAN}}$ such that $\Sigma \subset \Sigma'$ it is the case that $\Sigma' \vdash_{\mathsf{LAN}} \bot$.*

In what follows, we make use of the standard properties of MCSs. See Section 2.2 for all the proofs. We use these properties implicitly throughout the section.

Lemma 4.2 (Properties of MCSs)**.** *Let $\Sigma \subseteq \mathcal{L}^{\mathsf{LAN}}$ be an* LAN*-MCS and $\varphi, \psi \in \mathcal{L}^{\mathsf{LAN}}$:*

- *$\Sigma \vdash_{\mathsf{LAN}} \varphi$ iff $\varphi \in \Sigma$;*
- *$\varphi \in \Sigma$ iff $\neg\varphi \notin \Sigma$;*
- *$\varphi \wedge \psi \in \Sigma$ iff $\varphi \in \Sigma$ and $\psi \in \Sigma$.*

Adopting Lindenbaum's Lemma to the context of LAN, we know that every LAN-CS can be extended to an LAN-MCS.

Lemma 4.3 (Lindenbaum's Lemma)**.** *Let $\Sigma \subseteq \mathcal{L}^{\mathsf{LAN}}$ be an* LAN*-CS: there is an* LAN*-MCS $\Sigma' \subseteq \mathcal{L}^{\mathsf{LAN}}$ such that $\Sigma \subseteq \Sigma'$.*

Definition 4.9 (Canonical model for LAN)**.** *We define the* canonical model *to be the tuple $\mathfrak{M}^c := \langle W^c, \{W^c_{\mathsf{d}^\delta_i} \mid \mathsf{d}^\delta_i \in \mathcal{L}^{\mathsf{LAN}}\}, \{W^c_{\mathsf{v}_i} \mid i \in \mathsf{Agents}\}, \mathcal{R}^c_{\boxed{\mathsf{S}}}, \mathcal{R}^c_{\boxed{\mathsf{A}}}, V^c\rangle$ such that:*

- *$W^c := \{\Sigma \subset \mathcal{L}^{\mathsf{LAN}} \mid \Sigma$ is a* LAN*-MCS$\}$;*
- *For all $\mathsf{d}^\delta_i \in \mathsf{Wit}$, $W^c_{\mathsf{d}^\delta_i} := \{\Sigma \in W^c \mid \mathsf{d}^\delta_i \in \Sigma\}$;*

- *For all* $\mathtt{v}_i \in \mathsf{Vio}$, $W^c_{\mathtt{v}_i} := \{\Sigma \in W^c \mid \mathtt{v}_i \in \Sigma\}$;
- *For all* $\Sigma \in W^c$, $\mathcal{R}^c_{\boxed{\scriptstyle S}}(\Sigma) := \{\Theta \in W^c \mid$ *for all* $\boxed{\scriptstyle S}\varphi \in \Sigma,\ \varphi \in \Theta\}$;
- *For all* $\Sigma \in W^c$, $\mathcal{R}^c_{\boxed{\scriptstyle A}}(\Sigma) := \{\Theta \in W^c \mid$ *for all* $\boxed{\scriptstyle A}\varphi \in \Sigma,\ \varphi \in \Theta\}$;
- V^c *is a valuation function such that for all* $\chi \in \mathsf{Atoms} \cup \mathsf{Wit} \cup \mathsf{Vio}$, $V^c(\chi) := \{\Sigma \in W^c \mid \chi \in \Sigma\}$.

The semantic evaluation of $\mathcal{L}^{\mathsf{LAN}}$ *formulae on* $\mathfrak{M}^c$ *is defined as in Definition 4.7.*

The canonical model possesses the usual properties. The proofs of Lemma 4.4 and 4.5 are similar to those in Section 2.2.

Lemma 4.4 (Existence Lemma $\boxed{\scriptstyle S}$ and $\boxed{\scriptstyle A}$)**.** *For each* $\Sigma \in W^c$ *of* $\mathfrak{M}^c$ *the following holds:*

- *If* $\Diamond\!\!\!\!\!{\scriptstyle S}\ \varphi \in \Sigma$, *then there is a* $\Theta \in W^c$ *such that* $\varphi \in \Theta$ *and* $\Theta \in \mathcal{R}_{\boxed{\scriptstyle S}}(\Sigma)$;
- *If* $\Diamond\!\!\!\!\!{\scriptstyle A}\ \varphi \in \Sigma$, *then there is a* $\Theta \in W^c$ *such that* $\varphi \in \Theta$ *and* $\Theta \in \mathcal{R}_{\boxed{\scriptstyle A}}(\Sigma)$.

Corollary 4.2. *For any world* $\Sigma \in W^c$ *of* $\mathfrak{M}^c$ *the following holds:*

- *If for all* $\Theta \in \mathcal{R}_{\boxed{\scriptstyle S}}(\Sigma), \varphi \in \Theta$, *then* $\boxed{\scriptstyle S}\varphi \in \Sigma$;
- *If for all* $\Theta \in \mathcal{R}_{\boxed{\scriptstyle A}}(\Sigma), \varphi \in \Theta$, *then* $\boxed{\scriptstyle A}\varphi \in \Sigma$.

The following lemma shows that the defined model is canonical for LAN, i.e., each LAN-MCS is satisfiable on this model.

Lemma 4.5 (Truth Lemma)**.** *For any* $\varphi \in \mathcal{L}^{\mathsf{LAN}}$ *and* $\Sigma \in W^c$: $\mathfrak{M}^c, \Sigma \models \varphi$ *iff* $\varphi \in \Sigma$.

We show that the canonical model satisfies the desired behavior of action types.

Lemma 4.6. *For any* $\Delta \in \mathcal{L}^{\mathsf{Act}}$, *let* $W^c_{t(\Delta)}$ *be defined as in Definition 4.5. For each* $\Sigma \in W^c$ *the following holds:* $\mathfrak{M}^c, \Sigma \models t(\Delta)$ *iff* $\Sigma \in W_{t(\Delta)}$.

Proof. The proof is similar to Lemma 4.1 on page 138. QED

Last, we demonstrate that the defined canonical model is, in fact, a LAN-model.

Lemma 4.7 (Canonical LAN-model)**.** *The canonical model* $\mathfrak{M}^c$ *is a* LAN*-model.*

Proof. It can be easily observed that W^c and V^c (for $\chi \in \mathsf{Atoms} \cup \mathsf{Wit} \cup \mathsf{Vio}$) are well-defined (cf. Lemma 4.6). We show that $\mathfrak{M}^c$ satisfies the properties **R1**–**R6**.

R1 For each $\mathsf{d}^\delta_i \in \mathsf{Wit}$, $W^c_{\mathsf{d}^\delta_i} \subseteq W^c$ follows directly from the definition of $W^c_{\mathsf{d}^\delta_i}$.

R2 For each $\mathsf{v}_i \in \mathsf{Vio}$, $W^c_{\mathsf{v}_i} \subseteq W^c$. Similar to **R1**.

R3 Take arbitrary $\Sigma, \Theta, \Pi \in W^c$ and assume $\Theta \in \mathcal{R}^c_{\boxed{\mathsf{A}}}(\Sigma)$ and $\Pi \in \mathcal{R}^c_{\boxed{\mathsf{A}}}(\Sigma)$. Suppose towards a contradiction that $\Theta \neq \Pi$. Then, there is a $\varphi \in \Theta$ such that $\varphi \notin \Pi$, and so, $\neg\varphi \in \Pi$. By Lemma 4.5, $\mathfrak{M}^c, \Theta \models \varphi$ and $\mathfrak{M}^c, \Pi \models \neg\varphi$. Consequently, $\mathfrak{M}^c, \Sigma \models \langle\mathsf{A}\rangle\varphi$ and $\mathfrak{M}^c, \Sigma \models \langle\mathsf{A}\rangle\neg\varphi$ and so $\langle\mathsf{A}\rangle\varphi, \langle\mathsf{A}\rangle\neg\varphi \in \Sigma$. Since Σ is LAN-MCS, we have $\langle\mathsf{A}\rangle\varphi \rightarrow \boxed{\mathsf{A}}\varphi \in \Sigma$ (axiom $\mathsf{A3}$) and so $\boxed{\mathsf{A}}\varphi \in \Sigma$. However, by duality, this means $\neg\langle\mathsf{A}\rangle\neg\varphi \in \Sigma$. Contradiction.

R4 Take arbitrary $\Sigma, \Theta \in W^c$ and assume $\Theta \in \mathcal{R}^c_{\boxed{\mathsf{A}}}(\Sigma)$. Take an arbitrary $\boxed{\mathsf{S}}\varphi \in \Sigma$. Since Σ is a LAN-MCS, $\boxed{\mathsf{S}}\varphi \rightarrow \boxed{\mathsf{A}}\varphi \in \Sigma$ (axiom $\mathsf{A4}$) and so $\boxed{\mathsf{A}}\varphi \in \Sigma$. Since $\Theta \in \mathcal{R}^c_{\boxed{\mathsf{A}}}(\Sigma)$ by definition $\varphi \in \Theta$ and since $\boxed{\mathsf{S}}\varphi$ was arbitrary, $\Theta \in \mathcal{R}^c_{\boxed{\mathsf{S}}}(\Sigma)$.

R5 Assume $1, \dots, n \in \mathsf{Agents}$, $\Delta_1, \dots, \Gamma_n \in \mathcal{L}^{\mathsf{Act}}$, and $\Sigma, \Theta_1, \dots, \Theta_n \in \mathcal{R}^c_{\boxed{\mathsf{S}}}(\Sigma)$ and $\Theta_1 \in W^c_{t(\Delta_1)}, \dots, \Theta_n \in W^c_{t(\Gamma_n)}$. By Lemma 4.6, we have $\mathfrak{M}^c, \Theta_1 \models t(\Delta_1), \dots, \mathfrak{M}^c, \Theta_n \models t(\Gamma_n)$ and so $\mathfrak{M}^c, \Sigma \models \langle\mathsf{S}\rangle t(\Delta_1) \wedge \dots \wedge \langle\mathsf{S}\rangle t(\Gamma_n)$. By Lemma 4.5, $\langle\mathsf{S}\rangle t(\Delta_1) \wedge \dots \wedge \langle\mathsf{S}\rangle t(\Gamma_n) \in \Sigma$. By the fact that Σ is a LAN-MCS, we have $(\langle\mathsf{S}\rangle t(\Delta_1) \wedge \dots \wedge \langle\mathsf{S}\rangle t(\Gamma_n)) \rightarrow \langle\mathsf{S}\rangle(t(\Delta_1) \wedge \dots \wedge t(\Gamma_n)) \in \Sigma$ (axiom $\mathsf{A5}$) and so:

$$(\dagger) \quad \langle\mathsf{S}\rangle(t(\Delta_1) \wedge \dots \wedge t(\Gamma_n)) \in \Sigma$$

We show there exists a $\Pi \in W^c$ such that $\Pi \in \mathcal{R}^c_{\boxed{\mathsf{S}}}(\Sigma)$ and $\Pi \in W^c_{t(\Delta_1)} \cap \dots \cap W^c_{t(\Gamma_n)}$. Let,

$$\Pi' = \{\varphi \mid \boxed{\mathsf{S}}\varphi \in \Sigma\} \cup \{(t(\Delta_1) \wedge \dots \wedge t(\Gamma_n))\}$$

and suppose towards a contradiction that Π' is not LAN-consistent. Then, $\vdash_{\mathsf{LAN}} (\varphi_1 \wedge \dots \wedge \varphi_m) \rightarrow \neg(t(\Delta_1) \wedge \dots \wedge t(\Gamma_n))$ for some $\varphi_1, \dots, \varphi_m \in \{\varphi \mid \boxed{\mathsf{S}}\varphi \in \Sigma\}$. By the normality of $\boxed{\mathsf{S}}$, we obtain $\vdash_{\mathsf{LAN}} \boxed{\mathsf{S}}(\varphi_1 \wedge \dots \wedge \varphi_m) \rightarrow \boxed{\mathsf{S}}\neg(t(\Delta_1) \wedge \dots \wedge t(\Gamma_n))$. Again, by normality, $\vdash_{\mathsf{LAN}} (\boxed{\mathsf{S}}\varphi_1 \wedge \dots \wedge \boxed{\mathsf{S}}\varphi_m) \rightarrow \neg\langle\mathsf{S}\rangle(t(\Delta_1) \wedge \dots \wedge t(\Gamma_n))$. Since $\boxed{\mathsf{S}}\varphi_i \in \Sigma$ for $1 \leq i \leq m$ we have $\neg\langle\mathsf{S}\rangle(t(\Delta_1) \wedge \dots \wedge t(\Gamma_n)) \in \Sigma$ which contradicts $(\dagger)$. Hence, Π' is LAN-consistent. Let Π be the LAN-MCS extending Π'. By construction of Π' we know $\Pi \in \mathcal{R}^c_{\boxed{\mathsf{S}}}(\Sigma)$. Furthermore, since $t(\Delta_1) \wedge \dots \wedge t(\Gamma_n) \in \Pi$, by Lemma 4.5 we have $\mathfrak{M}^c, \Pi \models t(\Delta_1) \wedge \dots \wedge t(\Gamma_n)$. By Lemma 4.6, $\Pi \in W^c_{t(\Delta_1)}, \dots, \Pi \in W^c_{t(\Gamma_n)}$ and so $\Pi \in W^c_{t(\Delta_1)} \cap \dots \cap W^c_{t(\Gamma_n)}$.

R6 Take an arbitrary $\Sigma \in W^c$ and $i \in \mathsf{Agents}$. Suppose there is a $\Theta \in W^c$ such that $\Theta \in \mathcal{R}^c_{\boxed{\mathsf{S}}}(\Sigma)$ and $\Theta \in W^c_{\mathsf{v}_i}$. Hence, $\mathsf{v}_i \in \Theta$, which implies $\langle\mathsf{S}\rangle\mathsf{v}_i \in \Sigma$. Since Σ is a LAN-MCS we know $\langle\mathsf{S}\rangle\mathsf{v}_i \rightarrow \langle\mathsf{S}\rangle\neg\mathsf{v}_i \in \Sigma$ (axiom $\mathsf{A6}$) and so,

$$(\dagger) \quad \langle\mathsf{S}\rangle\neg\mathsf{v}_i \in \Sigma.$$

We show there exists a $\Sigma \in \mathcal{R}^c_{\boxed{\mathsf{S}}}(\Sigma)$ such that $\Sigma \in W^c \setminus W^c_{\mathsf{v}_i}$. Let,

$$\Sigma' = \{\varphi \mid \boxed{\mathsf{S}}\varphi \in \Sigma\} \cup \{\neg\mathsf{v}_i\}$$

and suppose towards a contradiction that Σ' is not LAN-consistent. Hence, $\vdash_{\mathsf{LAN}} (\varphi_1 \wedge \dots \wedge \varphi_n) \rightarrow \mathsf{v}_i$ for some $\varphi_1, \dots, \varphi_n \in \{\varphi \mid \boxed{\mathsf{S}}\varphi \in \Sigma\}$. By the normality of $\boxed{\mathsf{S}}$, $\vdash_{\mathsf{LAN}}$

$\boxed{\mathsf{S}}(\varphi_1 \wedge \cdots \wedge \varphi_n) \to \boxed{\mathsf{S}}\mathtt{v}_i$, which by duality means $\vdash_{\mathsf{LAN}} \boxed{\mathsf{S}}(\varphi_1 \wedge \cdots \wedge \varphi_n) \to \neg\langle\mathsf{S}\rangle\neg\mathtt{v}_i$. Since $\boxed{\mathsf{S}}(\varphi_1 \wedge \cdots \wedge \varphi_n) \in \Sigma$ by assumption, we have $\neg\langle\mathsf{S}\rangle\neg\mathtt{v}_i \in \Sigma$. Contradiction with (†) and the fact that Σ is a LAN-MCS. Hence, Σ' is LAN-consistent. Let Σ be the LAN-MCS extending Σ'. By construction of Σ we know $\Sigma \in \mathcal{R}^c_{\boxed{\mathsf{S}}}(\Sigma)$ and since $\neg\mathtt{v}_i \in \Sigma$, we have $\Sigma \notin W^c_{\mathtt{v}_i}$ which implies $\Sigma \in W^c \setminus W^c_{\mathtt{v}_i}$. QED

Theorem 4.2 (Strong Completeness of LAN)**.** *For any formula $\varphi \in \mathcal{L}^{\mathsf{LAN}}$, and any $\Theta \subseteq \mathcal{L}^{\mathsf{LAN}}$: if $\Theta \models_{\mathsf{LAN}} \varphi$, then $\Theta \vdash_{\mathsf{LAN}} \varphi$.*

Proof. The proof is similar to that of Theorem 2.2 on page 46 (Chapter 2). QED

Last, the satisfiability problem of LAN is decidable. This result was shown in (van Berkel et al., 2020) via proving the finite model property. We refer to the above work for its full proof.

Theorem 4.3 (Decidability of LAN)**.** *The satisfiability problem of* LAN *is decidable.*

4.4 Norms to Be, Norms to Do, and Norms of Instrumentality

The logic LAN allows us to reason about both actions and results. Accordingly, we distinguish three different types of normative statements: normative statements about (1) results, (2) actions, and (3) actions in relation to results. We refer to the first two categories as *norms to be* and *norms to do*, respectively, and to the third category as *norms of instrumentality*. The first two are generalizations of 'ought to be' and 'ought to do', e.g., (Castañeda, 1972). The latter category articulates which actions must or must not be employed when taken as instruments in relation to particular goals. In this section, we address Objective 1 and Objective 2 by providing logical formalizations of the three norm categories and determining their logical interdependencies.

Examples of the three categories are, respectively:

- "It ought to be (for agent *i*) that the registration for conference X is fulfilled";
- "Agent *i* ought to register for conference X before noon";
- "Agent *i* ought to register for conference X by means of filling out this form".

In this section, we demonstrate the expressive power of LAN by formally defining the abovementioned three categories. Subsequently, we use our formalization to investigate the dependencies between the different norm types. With this, we take the first step towards a formal analysis of norms of instrumentality. In the following section, we apply the attained definitions to formally analyze the example protocol (Section 4.1).

According to d'Altan et al. (1996), it is generally agreed upon that the categories of *ought to be* and *ought to do* cannot be entirely reduced to one another, even though they are related. In the introduction of this chapter, we discussed principle N1 of the protocol (page 126). We argued there that although a surgeon might be obliged to use a scalpel to ensure a required incision, it does not follow that she is obliged to use a scalpel independently of its intended purpose (some outcomes obtained by using scalpels could be prohibited), nor does it mean that she is obliged to make the incision by any means necessary (some means could be prohibited). Consequently, in the case of obligations *norms of instrumentality* are neither an instance of ought to be nor of ought to do obligations. The 'insider trading' example from the introduction (page 123) illustrates irreducibility for the case of prohibitions. Thus, all three norm categories must be considered as categories proper.

The question that arises is how these categories are related to one another. We stipulate three intuitive *desiderata* concerning the interdependencies between the three norm categories:

D1 If a result φ is prohibited, then φ will be prohibited regardless of the action used in obtaining it (i.e., prohibited given any action).

D2 If an action Δ is prohibited, then Δ's performance is prohibited irrespective of its outcome (i.e., prohibited given any outcome).

D3 If it is obligatory to perform a certain action Δ to obtain a particular result φ (instrumentality), then it is prohibited not to perform Δ, and it is prohibited not to bring about φ.

There are various ways to investigate the above desiderata. We distinguish between a *naive* approach to the three norm categories and a *deliberative* approach. The former deals with a simple formal representation of the three norm types. The latter concerns formalizations that take into account certain metaethical principles from the field of deontic logic. In what follows, we focus on the deliberative approach and point out along the way how to obtain the naive approach from these more involved definitions. Towards the end of this section, we discuss various other relations between the three categories.

We discuss two metaethical principles. The first is the principle of *no vacuous commands* which stipulates that norms are violable (see M1 below). The second adopted principle is that of ought implies can. We use the name *norm implies can* to refer to a generalized version of the principle, including prohibitions. We consider two interpretations of the term 'can'. First, we take 'can' to denote 'possible' (M2 below). Second, we interpret 'can' as an agent-dependent notion, referring to the agent's ability (M3 below). In what follows, we take norm implies can to express the provision that the agent can comply with the norm in question.[8] Henceforth, an obligation (or prohibition) that satisfies both M1 and M2 is called *deliberative* (cf. Chapter 3).

[8] We refer to (Governatori and Hashmi, 2015) for a discussion of compliance.

M1 It is possible to violate a norm: if X is prohibited (obligatory), then (the negation of) X is possible.

M2 It is possible to comply with a norm: if X is obligatory (prohibited), then (the negation of) X is possible.

M3 An agent is able to comply with a norm: if X is obligatory (prohibited), then the agent has the ability to guarantee (the negation of) X.

NB. Where X can be substituted for a result or an action.

At the end of this section, we argue that D1-D3 are not always M1-M3 compatible.

4.4.1 Norms to be

In what follows, we use the symbol $\mathcal{F}$ to refer to what is *forbidden* and use $\mathcal{O}$ to denote what is *obligatory*. Furthermore, we assume that the set Agents of agents and the set Act of action primitives are finite. Adapting Anderson's (1958) reduction to our deliberative approach, we formally define the first category of *norms to be* (i.e., *forbidden to be* and *ought to be*, respectively) as follows:

$$\begin{array}{lll}
\text{F1.} & \mathcal{F}_i(\varphi) & := & \boxed{s}(\varphi \to \mathsf{v}_i) \wedge \langle s\rangle\varphi \\
\text{O1.} & \mathcal{O}_i(\varphi) & := & \boxed{s}(\neg\varphi \to \mathsf{v}_i) \wedge \langle s\rangle\neg\varphi
\end{array}$$

We interpret $\mathcal{F}_i(\varphi)$ as "φ is forbidden to be the case for agent i, iff (i) every possible transition to φ would mean a norm violation for agent i and (ii) φ is possible". We read $\mathcal{O}_i(\varphi)$ as "φ ought to be the case for agent i, iff (i) every possible transition to $\neg\varphi$ would mean a norm violation for agent i and (ii) $\neg\varphi$ is possible". The first conjunct (i) of F1 and O1 corresponds to Anderson's reduction.[9] We refer to this clause as the *reduction clause*. The second conjunct (ii) captures that the norm can be violated. We refer to it as the *violation clause*.

Principle M1 is explicitly satisfied by definitions F1 and O1, cf. the violation clause. What is more, in LAN, these two definitions additionally satisfy M2. To see this point, suppose $\mathcal{F}_i(\varphi)$. By definition, $\langle s\rangle\varphi$ holds. Through basic LAN reasoning with the reduction clause, $\langle s\rangle\mathsf{v}_i$ can be inferred. By applying axiom A6, we obtain $\langle s\rangle\neg\mathsf{v}_i$. Last, by basic LAN reasoning with the reduction clause, we derive $\langle s\rangle\neg\varphi$. A similar argument can be given for O1. Hence, the following formula is LAN valid:

$$(\mathcal{F}_i(\varphi) \vee \mathcal{O}_i(\varphi)) \to (\langle s\rangle\varphi \wedge \langle s\rangle\neg\varphi)$$

In other words, definitions F1 and O1 express *deliberative* versions of norms to be: the norm can be both violated and complied with.

[9] To preserve a strict distinction between norm to be and norm to do, one may additionally stipulate that φ in F1 and O1 must be free of action constants from Wit.

We can extend the above definitions to capture *norms to be* under principle M3. We write $\mathcal{F}^*_i(\cdot)$ and $\mathcal{O}^*_i(\cdot)$ to denote this alternative formalization. Since we assumed that $\mathcal{L}^{\mathsf{Act}}$ is constructed over a finite number of action primitives from Act, there are only finitely many equivalence classes of action types $[\![\Delta_i]\!] := \{\Gamma_i \mid\ \models_{\mathsf{LAN}} t(\Gamma_i) \equiv t(\Delta_i)\}$ of equivalent actions. Let $[\![\mathcal{L}^{\mathsf{Act}}]\!]$ in F1* and O1* represent the set of all such equivalence classes.

$$\text{F1}^*. \qquad \mathcal{F}^*_i(\varphi) \quad := \quad \boxed{\mathsf{S}}(\varphi \to \mathrm{v}_i) \wedge \Diamond_{\mathsf{S}}\varphi \wedge \bigvee_{[\![\Delta_i]\!] \in [\![\mathcal{L}^{\mathsf{Act}}]\!]} (\boxed{\mathsf{S}}(t(\Delta_i) \to \neg\varphi) \wedge \Diamond_{\mathsf{S}} t(\Delta_i))$$

$$\text{O1}^*. \qquad \mathcal{O}^*_i(\varphi) \quad := \quad \boxed{\mathsf{S}}(\neg\varphi \to \mathrm{v}_i) \wedge \Diamond_{\mathsf{S}}\neg\varphi \wedge \bigvee_{[\![\Delta_i]\!] \in [\![\mathcal{L}^{\mathsf{Act}}]\!]} (\boxed{\mathsf{S}}(t(\Delta_i) \to \varphi) \wedge \Diamond_{\mathsf{S}} t(\Delta_i))$$

The deontic modalities $\mathcal{F}^*_i(\varphi)$ and $\mathcal{O}^*_i(\varphi)$ are similar to $\mathcal{F}_i(\varphi)$ and $\mathcal{O}_i(\varphi)$ in that they contain a reduction clause and a violation clause. They additionally contain a *norm implies ability* clause. This third clause expresses that (iii) "there exists an action available to the agent that is an instrument for complying with the norm, and the agent is able to perform that action" (cf. the 'could' operator in Section. 4.2). Observe that the third clause requires the action in question to be possible, thus excluding impossible actions such as $\delta_i \& \overline{\delta_i}$. Last, definitions F1* and O1* trivially satisfy principles M1 and M2 (by extending F1, respectively O1). The third clause explicitly satisfies M3.

Remark 4.5 (The Naive Approach)**.** *The* naive approach *to F1 and O1 is obtained by only considering the reduction clause (i). This holds true for all the formal definitions presented in this section.*

4.4.2 Norms to do

For the second category of *norms to do*, we adopt Meyer's (1988) reduction to the context of LAN. As for norms to be, we adopt a deliberative approach. *Forbidden to do* and *ought to do* are defined as follows:

$$\text{F2}. \qquad \mathcal{F}_i[\Delta] \quad := \quad \boxed{\mathsf{S}}(t(\Delta_i) \to \mathrm{v}_i) \wedge \Diamond_{\mathsf{S}} t(\Delta_i)$$

$$\text{O2}. \qquad \mathcal{O}_i[\Delta] \quad := \quad \boxed{\mathsf{S}}(\neg t(\Delta_i) \to \mathrm{v}_i) \wedge \Diamond_{\mathsf{S}} \neg t(\Delta_i)$$

We read $\mathcal{F}_i[\Delta]$ as "the performance of Δ is forbidden for agent i, iff (i) every possible performance of Δ would mean a norm violation for agent i and (ii) Δ can be performed by i". and we interpret $\mathcal{O}_i[\Delta]$ as "Δ ought to be performed by agent i, iff (i) every possible performance of $\overline{\Delta}$ would mean a norm violation for agent i and (ii) $\overline{\Delta}$ can be performed by i". The reduction clause (i) of F2 and O2 corresponds to Meyer's deontic reduction. Clause (ii) represents the violation clause. In passing, we point out that the reduction clause $\boxed{\mathsf{S}}(t(\Delta_i) \to \mathrm{v}_i)$ is interpretable in terms of instrumentality, i.e., "Δ is a v_i-instruments for agent i".

The above definitions of norms to do are deliberative, i.e., the following formula is LAN valid:

$$(\mathcal{F}_i[\Delta] \vee \mathcal{O}_i[\Delta]) \to (\Diamond\kern-0.6em{\scriptstyle S}\kern0.2em t(\Delta_i) \wedge \Diamond\kern-0.6em{\scriptstyle S}\kern0.2em \neg t(\Delta_i))$$

The reasoning is similar to the case of norms to be. We point out that the distinction between principles M2 and M3 breaks down for norms to do. Namely, in the case of $\mathcal{O}_i[\Delta]$ we take $\Diamond\kern-0.6em{\scriptstyle S}\kern0.2em t(\Delta_i)$ to express the idea that agent i has the ability to perform Δ because a successful performance of Δ by agent i is possible. The same applies to $\mathcal{F}_i[\Delta]$.

4.4.3 Norms of instrumentality

So far, the first two categories were obtained by extending their converged interpretation in the literature—see (d'Altan et al., 1996)—to a deliberative setting. How can we formally capture the third, novel category of *norms of instrumentality*? The above analysis suggests a definition containing at least a reduction clause and a violation clause. However, for norms of instrumentality, this does not suffice.

We start with *obligations* belonging to norms of instrumentality. First, we identify what it means for an agent to violate an obligation of the third category. If an agent i is obliged to employ Δ (as an instrument) to attaining φ, then i violates this obligation whenever *either* i does not perform Δ (independently of whether i produces φ) *or* i does not bring about φ (independently of whether i performs Δ). Second, recall that we take as *instruments* those actions suitable for serving a particular purpose. Hence, for an agent to be bound by such an obligation, we require that the prescribed action is, in fact, an *instrument* for bringing about the intended result. Based on the above two observations, we thus say that "an agent i is obliged to employ Δ as an instrument to obtaining φ iff (i) performing $\overline{\Delta}$ or bringing about $\neg\varphi$ would lead to a norm violation for agent i, (ii) such a norm violation is possible through $\neg\varphi$ or $\overline{\Delta}$, and (iii) the performance of Δ by i would ensure φ (i.e., Δ is a φ-instrument for i)". We formally define this obligation as follows:

$$\text{O3.} \qquad \mathcal{O}_i[\Delta](\varphi) \quad := \quad \boxed{\scriptstyle S}(\neg(t(\Delta_i) \wedge \varphi) \to \mathrm{v}_i) \wedge \Diamond\kern-0.6em{\scriptstyle S}\kern0.2em\neg(t(\Delta_i) \wedge \varphi) \wedge \boxed{\scriptstyle S}(t(\Delta_i) \to \varphi)$$

Notice that in the three conjuncts of definition O3, we identify (i) a reduction clause, (ii) a violation clause, and (iii) an instrumentality clause, respectively. O3 satisfies M1 by virtue of the second conjunct. That M2 and M3 are also satisfied can be straightforwardly shown: By applying basic LAN reasoning, the first and second conjunct imply $\Diamond\kern-0.6em{\scriptstyle S}\kern0.2em\mathrm{v}_i$. Together with axiom A6, we obtain $\Diamond\kern-0.6em{\scriptstyle S}\kern0.2em\neg\mathrm{v}_i$ which, by modus tollens on the first conjunct, implies $\Diamond\kern-0.6em{\scriptstyle S}\kern0.2em(t(\Delta_i) \wedge \varphi)$. Consequently, the following formula is LAN valid:

$$\mathcal{O}_i[\Delta](\varphi) \to [\Delta_i]^{could}\varphi$$

Can we similarly formalize prohibitions? Reconsider the 'insider trading' example from the introduction of this chapter (page 123): "it is prohibited to use non-public information as an instrument for attaining financial profit on the stock market". We say that an agent i violates this prohibition whenever i uses non-public information and consequently attains financial profit from it. Should we also say that i is only subject to this prohibition whenever she can guarantee a financial profit by using non-public information (as in the case of O3)? We believe that the answer is negative for prohibitions: we want to prohibit cases where i accidentally obtains financial profit on the stock market through using non-public information.[10] The resulting definition is expressed by F3.

$$\text{F3.} \qquad \mathcal{F}_i[\Delta](\varphi) \quad := \quad \boxed{\mathsf{S}}((t(\Delta_i) \wedge \varphi) \rightarrow \mathtt{v}_i) \wedge \Diamond_{\mathsf{S}}(t(\Delta_i) \wedge \varphi) \wedge (\Diamond_{\mathsf{S}}\neg t(\Delta_i) \vee \Diamond_{\mathsf{S}}\neg\varphi)$$

In F3, the first clause contains the reduction, the second clause expresses violability, and the third clause captures the possibility of complying with the prohibition, i.e., either not performing the prohibited action is possible or the prohibited outcome is avoidable. Consequently, F3 satisfies M1 and M2 by definition. That is, the following formula is LAN valid:

$$(\mathcal{O}_i[\Delta](\varphi) \vee \mathcal{F}_i[\Delta](\varphi)) \rightarrow (\Diamond_{\mathsf{S}}(t(\Delta_i) \wedge \varphi) \wedge \Diamond_{\mathsf{S}}\neg(t(\Delta_i) \wedge \varphi))$$

To account for M3, we require that the agent has the *ability* to comply with the prohibition. In that case, we say that "agent i is prohibited from obtaining φ by means of performing action Δ, iff (i) in every case in which Δ has been performed and φ has been successfully acquired, a norm violation has occurred, (ii) the prohibition can be violated, and (iii) either agent i has the ability to avoid performing Δ or there is an action at i's disposal that is an instrument for avoiding φ". Formally, this definition is expressed accordingly:

$$\text{F3}^*. \qquad \mathcal{F}^*_i[\Delta](\varphi) \quad := \quad \boxed{\mathsf{S}}((t(\Delta_i) \wedge \varphi) \rightarrow \mathtt{v}_i) \wedge \Diamond_{\mathsf{S}}(t(\Delta_i) \wedge \varphi) \wedge \theta$$

$$\text{where } \theta := \Diamond_{\mathsf{S}}\neg t(\Delta_i) \vee \bigvee_{[\![\Gamma_i]\!] \in [\![\mathcal{L}^{\mathsf{Act}}]\!]^*} (\boxed{\mathsf{S}}(t(\Gamma_i) \rightarrow \neg\varphi) \wedge \Diamond_{\mathsf{S}} t(\Gamma_i))$$

The first two conjuncts of F3* correspond to the reduction and violation clause, respectively. The third conjunct explicitly stipulates the agent's ability to comply with the command. It can be easily shown that F3* implies F3. That is, the formula

$$\mathcal{F}^*_i[\Delta](\varphi) \rightarrow \mathcal{F}_i[\Delta](\varphi)$$

is LAN valid. Consequently, $\mathcal{F}^*_i[\Delta](\varphi)$ satisfies M1 and M2 too. Moreover, the following formula is LAN valid:

$$\mathcal{F}^*_i[\Delta](\varphi) \rightarrow ([\overline{\Delta_i}]^{could}\top \vee \bigvee_{[\![\Gamma_i]\!] \in [\![\mathcal{L}^{\mathsf{Act}}]\!]^*} [\Gamma_i]^{could}\neg\varphi)$$

[10] An alternative definition—akin to O3—in which the agent is only subject to the prohibition whenever she can guarantee the result through performing the action in question is straightforwardly obtained.

In other words, given $\mathcal{F}^*_i[\Delta](\varphi)$, the agent either has the ability not to perform Δ or she has the ability to ensure that φ does not hold. We believe that the third conjunct in O3 and F3* is pivotal for *deliberative* norms of instrumentality: it ensures that the outcome of compliance is a proper consequence of the agent's conduct. Again, the naive approach to norms of instrumentality is obtained by merely adopting the first clause of definition O3, F3, and F3*, i.e., the reduction clause.

4.4.4 Relations Between the Three Norm Categories

We now discuss the interaction between the proposed formal definitions and the desiderata presented at the beginning of this section. We do this both with respect to the naive and the deliberative approach. Desiderata D1 and D2 are formalized as

$$\mathcal{F}'_i(\varphi) \to \mathcal{F}'_i[\Delta](\varphi)$$

respectively,

$$\mathcal{F}_i[\Delta] \to \mathcal{F}'_i[\Delta](\varphi)$$

where $\mathcal{F}'_i \in \{\mathcal{F}_i, \mathcal{F}^*_i\}$. First, we observe that the above three formulae are LAN-valid for the naive approach to the three deontic modalities (the distinction between $\mathcal{F}_i$ and $\mathcal{F}^*_i$ disappears for the naive approach). To see this point, observe that if at every moment at which φ holds, a violation v_i ensues, then a fortiori at every moment at which φ and $t(\Delta_i)$ hold, a violation v_i ensues (the same reasoning applies for the second formula). In contrast, a straightforward LAN-model can be constructed to show that these formulae are not valid for the deliberative approach. The reason is the violation clause in the formula $\mathcal{F}'_i[\Delta](\varphi)$ which requires that $\Diamond(t(\Delta_i) \wedge \varphi)$. Roughly, a prohibition against bringing about a result (action) does not imply that the result (action) must be avoided given any action (result) but only given every action (result) compatible with the result (action). Consequently, impossible combinations of actions and results cannot be forbidden because they are *inviolable* (M1). The above results are represented in Table 4.1-V1.

As stated in D3, when an agent i is obliged to ensure φ by means of performing Δ, we want to conclude that both the state of affairs $\neg\varphi$ and the performance of $\overline{\Delta}$ are prohibited for agent i. This desideratum is formally expressed by

$$\mathcal{O}_i[\Delta](\varphi) \to \mathcal{F}'_i(\neg\varphi)$$

and

$$\mathcal{O}_i[\Delta](\varphi) \to \mathcal{F}_i[\overline{\Delta}]$$

where $\mathcal{F}'_i \in \{\mathcal{F}_i, \mathcal{F}^*_i\}$. Both formulae are LAN valid for the naive approach. Concerning the deliberative approach, the first formula is not LAN valid. Here too, the main reason lies in the violation clause. Namely, the definition of $\mathcal{O}_i[\Delta](\varphi)$ requires that the obligation is violable. This is the case when $\overline{\Delta}$ can be performed or when $\neg\varphi$ can be obtained. The disjunction is insufficient for concluding that $\neg\varphi$ is obtainable, which is required by the violation clause of $\mathcal{F}_i(\neg\varphi)$. In contrast, the second formula is LAN valid for the

				Naive	Deliberative
V1.	$\mathcal{F}'_i(\varphi) \to \mathcal{F}'_i[\Delta](\varphi)$	and	$\mathcal{F}_i[\Delta] \to \mathcal{F}'_i[\Delta](\varphi)$	yes	no
V2.	$\mathcal{O}_i[\Delta](\varphi) \to \mathcal{F}'_i(\neg\varphi)$	and	$\mathcal{O}_i[\Delta](\varphi) \to \mathcal{F}_i[\overline{\Delta_i}]$	yes	no, resp. yes
V3.	$\mathcal{O}'_i(\varphi) \to \mathcal{F}'_i[\Delta](\neg\varphi)$	and	$\mathcal{O}_i[\Delta] \to \mathcal{F}'_i[\overline{\Delta}](\varphi)$	yes	no
V4.	$\mathcal{F}'_i(\varphi) \to \mathcal{O}_i[\Delta](\neg\varphi)$	and	$\mathcal{F}_i[\Delta] \to \mathcal{O}_i[\overline{\Delta}](\varphi)$	no	no
V5.	$\mathcal{F}'_i(\varphi) \equiv \mathcal{O}'_i(\neg\varphi)$	and	$\mathcal{F}_i[\Delta] \equiv \mathcal{O}_i[\overline{\Delta}]$	yes	yes
V6.	$\mathcal{O}_i[\Delta](\varphi) \to \mathcal{O}'_i(\varphi)$	and	$\mathcal{O}_i[\Delta](\varphi) \to \mathcal{O}_i[\Delta]$	yes	no, resp. yes
V7.	$\mathcal{O}_i[\Delta] \wedge \mathcal{O}'_i(\varphi) \to \mathcal{O}_i[\Delta](\varphi)$			yes	no
V8.	$\mathcal{F}'_i(\varphi) \wedge \mathcal{F}_i[\Delta] \to \mathcal{F}'_i[\Delta](\varphi)$			yes	no

Table 4.1: The table contains implications between various deontic formulae. The deontic formulae are based on definitions F1-F3, O1-O3, F1*, and O1*. Let $\mathcal{F}'_i \in \{\mathcal{F}_i, \mathcal{F}^*_i\}$ and $\mathcal{O}'_i \in \{\mathcal{O}_i, \mathcal{O}^*_i\}$. The penultimate column represents the naive approach. The last column represents the deliberative approach. We write 'yes' to indicate that the formula in question is LAN valid and 'no' if otherwise.

deliberative approach. To see this point, we make two observations: First, we know that if performing $\overline{\Delta}$ or attaining $\neg\varphi$ leads to a violation, then $\overline{\Delta}$ yields a violation. This yields the reduction clause of $\mathcal{F}_i[\overline{\Delta}]$. Second, $\mathcal{O}_i[\Delta](\varphi)$ requires that the obligation is violable, i.e., $\overline{\Delta}$ is performable or $\neg\varphi$ is attainable. In both cases, this implies the violation clause of $\mathcal{F}_i[\overline{\Delta}]$. For the first disjunct, this is trivial. For the second disjunct, it follows from the fact that Δ is a *φ-instrument*, which means that if $\neg\varphi$ attains, Δ was not performed. See Table 4.1-V2.

In Table 4.1, we present various implications between various deontic formulae that bear significance to the present analysis. The deontic formulae are based on definitions F1-F3, O1-O3, F1*, and O1*. The penultimate column represents the naive approach, i.e., assuming that the deontic formulae in question contain only the reduction clause (i). The last column represents the deliberative approach, where the deontic formulae are considered with all the defined clauses. The dependencies described by V4 and V5 are invariant to whether the naive or deliberative approach is adopted. In particular, V5 expresses that prohibition and obligation are interdefinable for norms to be and norms to do; cf. (d'Altan et al., 1996). Last, V7 and V8 show that, for the deliberative approach, even the combination of norms to be and norms to do is insufficient to yield a norm of instrumentality. The present analysis is the first step toward a thorough investigation of norms of instrumentality, and further analysis is left to future work.

4.5 Formal Examples

In what follows, we apply our formal machinery to the example in the introduction of this chapter (page 126). We formalize the protocol in LAN by using definitions F1-F3 and O1-O3 and apply it to two concrete situations where an agent must invoke the protocol to make a decision. Our formalization will demonstrate that the protocol is insufficient relative to its assumed aims, i.e., T1 and T2.

In what follows, $\texttt{we use this font}$ to denote states of affairs, e.g., the formula $\texttt{incis}$ denotes the proposition "the incision is made", and $\textsf{we use this font}$ to denote actions. e.g., the action type $\textsf{scalp}$ denotes the action "using a scalpel". For the formalization of the protocol, we take s and n to denote the agents 'surgeon' and 'nurse', respectively. The action language $\mathcal{L}^{\textsf{Act}}$ consists of the atoms $\textsf{scalp}$, $\textsf{leave}$, and $\textsf{call}$, respectively describing "using a scalpel", "leaving the operation room" and "calling the safety-emergency number". Let $\texttt{incis}$, $\texttt{operation}$, $\texttt{dire}$, $\texttt{health}$, $\texttt{safety\_nur}$, and $\texttt{safety\_sur}$ be propositional atoms denoting "the incision is made", "the situation is an operation", "the situation is dire", "the patient's health is promoted", "hygiene safety is promoted from the nurse's perspective" and "hygiene safety is promoted from the surgeon's perspective", respectively. Consider the following formalization of the protocol:

P1. $(\texttt{operation} \land \mathcal{O}_s(\texttt{incis})) \to \mathcal{O}_s[\textsf{scalp}](\texttt{incis})$;

P2. $(\texttt{operation} \land \neg\texttt{dire}) \to \mathcal{F}_n[\textsf{scalp}]$;

P3. $\mathcal{O}_s(\texttt{health}) \land \mathcal{O}_s(\texttt{safety\_nur})$ and $\mathcal{O}_n(\texttt{health}) \land \mathcal{O}_n(\texttt{safety\_sur})$;

E1. $\neg\texttt{safety\_nur} \to (\mathcal{O}_s[\textsf{leave}] \land \mathcal{O}_s[\textsf{call}])$ and $\neg\texttt{safety\_sur} \to (\mathcal{O}_n[\textsf{leave}] \land \mathcal{O}_n[\textsf{call}])$.

As an example of how to read the formulae above, we interpret P2 as: "if there is an operation and the situation is not dire, then the nurse is prohibited from using the scalpel (irrespective of its outcome)". We are currently interested in whether the protocol is consistent and whether it can provide agents with sufficient tools to solve normative issues (in situations relevant to our example).[11] Concerning the former, the model in Figure 4.4 demonstrates the consistency of the protocol by satisfying P1-P3 and E1. Regarding the latter, let us consider some possible situations.

Situation 1. In the operation room, Anna the head surgeon, and a nurse named Bill, are performing a tonsillectomy on a patient (i.e., the patient's tonsils are to be removed). Anna must make a final highly demanding dissection involving both hands when she

[11] The logic LAN adopts classical logic and satisfies the principle of explosion: "from a contradiction, anything follows". Hence, in the context of LAN, an inconsistent protocol is undesirable because we would not be able to draw any meaningful conclusion from the protocol, namely, anything would follow. In paraconsistent and non-monotonic logics that do not satisfy this principle (such as the logics considered in Chapters 6 and 7), one can investigate which meaningful conclusions can be drawn from an inconsistent protocol.

realizes that another crucial incision has to be made using the harmonic scalpel (a scalpel that simultaneously cauterizes tissue). Since Anna is preoccupied and unable to do it, she appeals in this dire situation to Bill, asking whether he could make the other necessary incision with the harmonic scalpel, thus ensuring the patient's health. The situation is formalized accordingly:

(i) $\texttt{operation} \land \texttt{dire} \land [\overline{\mathsf{scalp}}_s]^{will}\top$;

(ii) $[\mathsf{scalp}_n]^{would}\texttt{incis}$;

(iii) $[\overline{\mathsf{scalp}}_n]^{would}\neg\texttt{health}$;

(iv) $\Box(\texttt{incis} \to \texttt{health})$.

Bill is aware of the new protocol: he knows he is not allowed to use scalpels in regular situations but remembers his duty to the patient's health. What should Bill do? The protocol tells Bill that he is obliged to promote the patient's health (i.e., $\mathcal{O}_n(\texttt{health})$, follows from P3). Since the surgical situation is dire (i), principle P2 does not apply. Moreover, since using the scalpel to make the incision is Bill's only way to promote the patient's health—by (ii)-(iv)—Bill is obliged to make the incision with the scalpel. That is, the following is LAN valid:

$$((\mathrm{i}) \land (\mathrm{ii}) \land (\mathrm{iii}) \land (\mathrm{iv}) \land \mathrm{P1} \land \mathrm{P2} \land \mathrm{P3} \land \mathrm{E1}) \to \mathcal{O}_n[\mathsf{scalp}](\texttt{incis})$$

Consequently, Bill is not prohibited from using the scalpel (i.e., $\neg\mathcal{F}_n[\mathsf{scalp}]$ follows from definition O3, LAN reasoning and V5 of Table 4.1).

Furthermore, to see whether Bill complies with the protocol when he *actually* brings about the incision with the scalpel, that is,

(v) $[\mathsf{scalp}_n]^{will}\texttt{incis}$

Consider the LAN-model in Figure 4.4. The model shows that Bill's behavior (v), together with the formalized protocol P1-P3 and E1 and the present situation (i)-(iv), can be consistently represented together with Bill's actual norm compliance, that is,

(vi) $\Diamond_{A}\neg\mathtt{v}_n$

In other words, (i)-(vi), P1-P3, and E1 are LAN-consistent, i.e., satisfiable on a LAN-model. For that reason, Bill's decision to make the incision using the scalpel preserves the state of compliance. Nevertheless, as expected, it can still be the case that a violation ensues due to some other action of Bill's. For instance, if Bill actually decides to *not* use the scalpel, a norm violation will be inevitable. That is, the following is valid:

$$((\mathrm{i}) \land (\mathrm{ii}) \land (\mathrm{iii}) \land (\mathrm{iv}) \land \mathrm{P1} \land \mathrm{P2} \land \mathrm{P3} \land \mathrm{E1} \land [\overline{\mathsf{scalp}}_n]^{will}\top) \to [\overline{\mathsf{scalp}}_n]^{will}\mathtt{v}_n$$

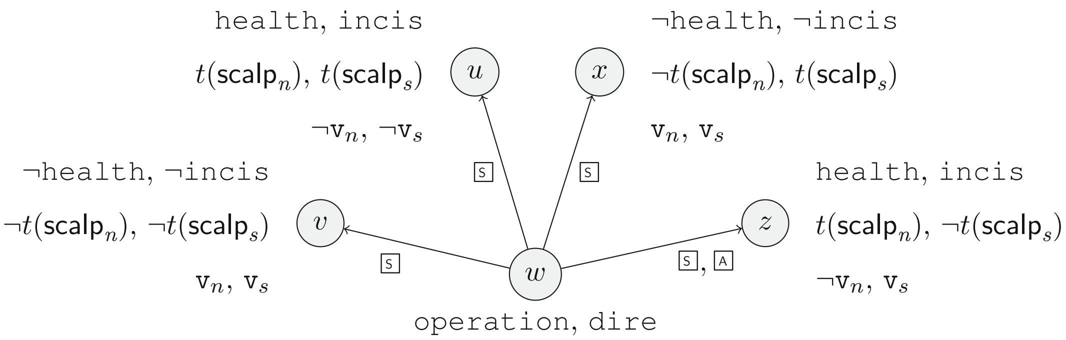

health, incis
$t(\mathsf{scalp}_n)$, $t(\mathsf{scalp}_s)$
$\neg\mathtt{v}_n$, $\neg\mathtt{v}_s$
$\neg$health, $\neg$incis
$\neg t(\mathsf{scalp}_n)$, $t(\mathsf{scalp}_s)$
$\mathtt{v}_n$, $\mathtt{v}_s$
$\neg$health, $\neg$incis
$\neg t(\mathsf{scalp}_n)$, $\neg t(\mathsf{scalp}_s)$
$\mathtt{v}_n$, $\mathtt{v}_s$
health, incis
$t(\mathsf{scalp}_n)$, $\neg t(\mathsf{scalp}_s)$
$\neg\mathtt{v}_n$, $\mathtt{v}_s$
u
x
v
z
w
S
S
S
S, A
operation, dire

Figure 4.4: A LAN-model satisfying P1-P3, E1 and (i)-(v). The model shows the consistency of the protocol and represents Bill's actual and compliant behavior in situation 1.

Situation 2. Let us continue the above example. Right before Bill performs the procedure involving the scalpel, Bill accidentally hits his own arm with the harmonic scalpel and inflicts a painful wound. Since Bill has now violated his obligation (P3) to preserve the required hygiene safety, Bill and Anna know that he is obliged (E1) to immediately leave the operation room and call the safety-emergency number for assistance. However, Anna observes that the necessary incision still has to be made to secure the agent's health. Hence, she concludes that Bill must stay and assist her immediately without further ado. The situation is formalized accordingly:

(vii) $\neg\mathtt{safety\_nur}$

(viii) $[\mathsf{leave}_n]^{would}\neg\mathtt{health}$

First, we observe that given E1 and (vii), Bill is obliged to leave (i.e., $\mathcal{O}_n[\mathsf{leave}]$). However, through (viii), the act of leaving would imply that Bill violates his obligation to preserve the patient's health (i.e., $\mathcal{O}_n(\mathtt{health})$). The current situation and the formalized protocol are inconsistent. Namely, (vii)-(viii), together with P1-P3 and E1, imply that Bill is obliged to leave and not to leave (i.e., $\mathcal{O}_n[\mathsf{leave}\&\overline{\mathsf{leave}}]$). The inconsistency depends on the assumption T1 (cf. A6 of Definition 4.6), which is the committee's assumption that there is a way out to every possible dilemma.

The primary purpose of this section was to illustrate the expressivity of LAN and the logical behavior of the three norm categories interacting. As a final remark, we observe that the source of the conflict in the second situation relates to contrary-to-duty (CTD) reasoning. Principle E1 is a contrary-to-duty obligation, and the above formalization suffers from a similar detachment problem as the formalization of Chisholm's (1963) CTD paradox in Standard Deontic Logic. We refer to Chapter 1 for an introduction. Like CTD obligations, E1 comes into force whenever the initial obligation in P3 is violated. The purpose of such an obligation is, then, to (partially) *restore* compliance with the norm system (e.g, Governatori and Rotolo, 2006; Governatori and Hashmi, 2015). Dynamic deontic logics, such as the one introduced by Meyer (1988), deal with CTD reasoning by giving such scenarios a temporal, action reading. Problems are then avoided since the primary obligation (cf. P3), and secondary obligation (cf. E1) occur at different moments in time. See (Prakken and Sergot, 1996) for challenges concerning temporal solutions to CTD reasoning. In Chapter 5, we extensively discuss the reductionist approach (adopted in this chapter) in relation to CTD reasoning.

4.6 Formal Notions of Instrumentality

So far, we have adopted a basic notion of moment-dependent instrumentality: $[\Delta_i]\varphi := \boxed{\mathsf{s}}(t(\Delta_i) \rightarrow \varphi)$. The definition sufficed for the analysis of norms of instrumentality. In other settings, more involved notions may be required. For instance, instrumentality—or means-end—statements play a central role in practical reasoning and planning: agents use such statements to deliberate about which action to perform to achieve a particular goal (Audi, 1989; Clarke, 1987).[12] Consider the following practical inference:

P1 I want to be on time for the rehearsal;

P2 I can only be on time if I take the A-train;

C1 Therefore, I must take the A-train.

Premise P2 expresses the means that enables the agent to reach the goal of "being on time at the rehearsal" (P1). That "taking the A-train" is the *only* means for achieving this goal *necessitates* the agent to perform the action of "taking the A-train" (C1).[13] For practical reasoning, it becomes important to *assess* statements such as P2. This means answering questions such as:

How are instrumentality judgments acquired by an agent?

How can we determine whether an instrument serves a purpose well?

In (van Berkel et al., 2022b), we discuss different ways in which such relations can be obtained, compared, and assessed. For the sake of completion, we informally discuss some observations made there and provide formal definitions of some more involved instrumentality statements. The account provided in (van Berkel et al., 2022b) is inspired by observations made by von Wright (1963a,1972b).

[12] In Section 4.7 we discuss the related Belief-Desire-Intention systems (Rao and Georgeff, 1995).

[13] See (Hare, 1971; Clarke, 1987; Walton, 2007) for practical reasoning with sufficient means. For von Wright's account of practical inference, see (1963a,1972b).

4.6.1 Assessing instrumentality through past experience

A central component for assessing instrumentality statements is an agent's *past experience* with performing the action in question. It suffices for an agent to consult her past experience without any additional knowledge of natural causality. This makes past experience particularly suitable as a guide in deliberation. We refer to this *temporal component* as the historical witness of an instrument's suitability.[14]

Everyday life deliberations—e.g., concerning a plan of action—are often based on statements such as:

(i) "it has worked before";

(ii) "so far, it has not disappointed me";

(iii) "well, thus far, it worked better than any of the alternatives".

The first remark exemplifies a minimum criterion for instrumentality: (i) the action *has* served the purpose at least once, and, for that reason, it *can* serve the purpose as an instrument. Criterion (i) functions as a lower bound on the instrument's suitability, thus identifying *potential instruments*. In the second remark, we recognize an upper bound, that is, a maximum criterion: (ii) there have been applications of the instrument, and these applications *have always* served the purpose. Criterion (ii) is referred to as instrumental *excellence*. In the last remark, we identify a *comparative* approach to instrumental goodness: (iii) the action is suitable in *comparison* to alternative actions. In what follows, we discuss (i) and (ii) and refer to (van Berkel et al., 2022b) for a formal analysis of definitions of comparative instrumentality. [15]

We obtain the following two definitions of instruments:

Definition 4.10. *instruments (with respect to past experience)*

(1) potential φ-instrument: An action-type Δ is a φ-instrument for agent i at moment w if and only if (i) Δ has led to φ at least once in the past.

(2) excellent φ-instruments: An action-type Δ is an excellent φ-instrument for agent i at moment w if and only if (i) Δ is a φ-instrument and (ii) Δ has always led to φ in the past.

[14] Von Wright (1972b, Ch.2) also discusses the assessment of instruments by investigating, what he calls, their "good-making properties". For instance, the sharpness of a knife enables us to objectively determine whether a knife serves the purpose of, say, cutting vegetables well. Sharpness is, in this respect, a good-making property. Von Wright emphasizes that although properties, such as 'sharp', may have vague meanings, the comparative 'sharper' has an objective empirical ordering. The ordering provides a logical and empirical method for determining which knives serve the purpose best (1972, p. 25). We refer to (van Berkel et al., 2022b) for an extensive discussion.

[15] For instance, Δ is a *better* φ-instrument for agent i at moment w than actions $\Gamma_1, \ldots, \Gamma_n$ iff (i) Δ is a φ-instrument and (ii) in the past Δ led to φ more frequently than the other φ-instruments $\Gamma_1, \ldots, \Gamma_n$.

Since we define instruments in relation to past performance—where experience functions as a historical *witness*—such a qualification is strictly context- and agent-dependent.

When we say that an action is a suitable instrument for a particular purpose, we do not refer to its causal physical qualities *per se*. We take the notion of an instrument as a practical one and keep it distinct from the idea of a cause, as employed in analyzing physical connections between things or events. In particular, we see past experience as a fruitful approach to agentive reasoning since it is a source of knowledge accessible to the agent at any given time (independent of the underlying causal connections). We note here that von Wright (1972b) does provide an analysis of causal connections concerning instrumentality. We briefly discuss the main idea. Consider a situation where an agent intends to open a parcel with a knife. The sharpness of each available knife determines which knife is the most suitable instrument. In this scenario, 'sharpness' determines the causal link between using a knife and cutting. The agent can, subsequently, order all available knives according to their sharpness to determine which are sharper and, thus, better instruments (cf. footnote 14 on page 156). We leave the analysis of instrumentality in the context of causal processes for future work.

4.6.2 Expectations: The Other Temporal Component

The past serves as a fruitful source for identifying instrumentality relations. Often such judgments are established via *inductive arguments*. The resulting generalizations are inherently defeasible because future information may falsify earlier judgments. For instance, "before cars, horses may have been the best means of private transport". Furthermore, In forming instrumentality judgments, an individual agent can often not collect all relevant past cases that would settle the issue. Nevertheless, using instrumentality statements, an agent can generalize past experience and project it onto the future in the form *expectations*. Expectations capture what von Wright calls the *conjectural element* in instrumentality judgments (1972b). An expectation concerning instruments is a projection of the past onto the nearby future by an explicitly expected continuation of the action serving the intended purpose. Furthermore, expectations explain how an agent's deliberation may be mistaken: she may have expected some other future moments to be possible.[16]

Definition 4.11. *INSTRUMENTS (WITH RESPECT TO PAST AND FUTURE)*

(1) POTENTIAL φ-INSTRUMENTS: *An action-type Δ is a φ-instrument for agent i at moment w if and only if (i) Δ has led to φ at least once in the past and (ii) i expects that Δ will lead to φ in the immediate future of w.*

(2) EXCELLENT φ INSTRUMENTS: *An action-type Δ is an excellent φ-instrument for agent i at moment w if and only if (i) Δ is a φ-instrument, (ii) Δ has always led to φ in the past and (iii) i expects that Δ will lead to φ in the immediate future of w.*

[16] Expectations must not be confused with notions of (incomplete) knowledge. An agent can have expectations about the future independently of the her knowledge of these expected future moments.

Instrumentality judgments are often established by inductive arguments. As famously noted by David Hume, inductive arguments are affected by a fundamental problem: how are we justified in making inferences from an observed connection in the past to instances of that connection of which we have no experience? Von Wright (1957) investigates the problem of induction by dividing it into two sub-questions: (a) How can we demonstrate that the generalizations we make about experienced cases are correct, and (b) how can we demonstrate that such generalizations are reliable for making predictions? Von Wright's division is temporal: the first question deals with the past, and the second with the future. Regarding generalizations extending to the past, one can theoretically acquire universally objective judgments by collecting all past instances of the object under generalization. However, when it comes to predictions, the problem of induction truly shows itself: "Scarcely anybody would pretend that predictions, even when based upon the safest inductions, might not fail sometimes" (von Wright, 1957, p. 51). Consequently, generalizations are inherently defeasible. Von Wright's account of instrumentality judgments (implicitly) incorporates the same temporal distinction between collecting past cases and extending past generalizations to the future through predictions. By restricting instrumentality judgments to (defeasible) expectations, the problem of induction does not prevent the agent from formulating instrumentality judgments that guide action in the immediate future.[17]

4.6.3 Some Formal Definitions

In order to formally illustrate the above analysis, we consider an extension of the language $\mathcal{L}^{\mathsf{LAN}}$. In (van Berkel et al., 2022b), we introduce a *Temporal Logic of Actions and Expectations*, which is a modification of LAN. The formalism adopts an explicitly temporal language in an indeterministic branching time setting. It consists of a back-ward looking modality $\blacksquare$ that enables reference to an agent's past experience and expectation constants e_i that model how an agent's past experience is projected onto the nearby future. In what follows, we consider this extension of $\mathcal{L}^{\mathsf{LAN}}$. We take $\blacksquare$ as the inverse of $\boxed{S}$ and read $\blacksquare\varphi$ as "everywhere in the immediate past of the present moment φ holds". We read e_i as "the most recent expectations of an agent i are fulfilled" (referring to the immediate predecessor).

Remark 4.6. *In (van Berkel et al., 2022b), we provide a sound and complete axiomatization of this modification of* LAN. *In particular, we axiomatize irreflexive treelike structures which branch towards the future and are linear with respect to the past. This means that* $\blacksquare$ *and* $\boxed{S}$ *are irreflexive modalities. Consequently, the formulae* $\blacklozenge\varphi$ *and* $\blacklozenge\blacklozenge\varphi$ *refer to two distinct moments in the past, namely, the immediate predecessor and the immediate predecessor after that. We use* $\blacklozenge^i$ *to refer to a concatenation of i-many* $\blacklozenge$. *Furthermore, since the past is linear,* $\blacklozenge\varphi \rightarrow \blacksquare\varphi$ *is an axiom of the logic. We adopt the*

[17] We point out that generalized statements establish norms in their own right. For instance, horses used to be the norm for fast personal transportation before the rise of the car. Such norms express which means are considered most appropriate for attaining certain ends. Since such norms are based on generalization, they are likewise defeasible.

axiom $\Diamond_{\mathrm{S}}\mathtt{e}_i \to \Diamond_{\mathrm{S}}\neg\mathtt{e}_i$ *to characterize the idea that if an agent expects a particular next moment to arise, there will be another next moment that the agent does not expect to arise, i.e., there is a limit to the agent's expectations. The axiom allows for situations in which the agent did not expect anything to happen. In the sequel, we omit formal details.*

We start with formalizing potential and excellent φ-instruments, which only consider the agent's past experience, i.e., (1) and (2) of Definition 4.10. Past experience is defined up to an interval of length n.

Potential φ-instrument for agent i (past)

$$[\Delta_i]_n^{p-instr}\varphi := \bigvee_{0\leq j<n} \blacklozenge^j(t(\Delta_i) \wedge \blacklozenge[\Delta_i]^{would}\varphi) \qquad (\mathtt{d}10)$$

The definition in (d10) is interpreted as "somewhere within the past interval of length n there is a moment at a distance of j units of time (at most $n-1$) that witnessed the successful performance of Δ by agent i and such that at the immediate predecessor at a distance of $j+i$ units in time (at most n) the performance of Δ by that agent would have guaranteed φ". Observe that as an immediate consequence, at the distance of j units of time, it is the case that φ holds. Furthermore, observe that for $\blacklozenge^j$, the value of j can be equal to 0. This means that the moment of evaluation is also included as a witness.

Excellent candidate instruments combine the above definition with the idea that *every* past performance of the relevant action type has led to the intended outcome.

Excellent φ-instrument for i (past)

$$[\Delta_i]_n^{exc-instr}\varphi := [\Delta_i]_n^{p-instr}\varphi \wedge \bigwedge_{1\leq j\leq n} \blacksquare^j[\Delta_i]^{would}\varphi \qquad (\mathtt{d}11)$$

That is, (d11) expresses that "the action-type Δ has proved to be a candidate instrument for φ for i at least once in the interval, and every other performance of Δ by i within the interval would have also guaranteed φ".

In order to incorporate expectations, we introduce two refined notions of the agentive operators *would* (d1) and *could* (d2):

Expected Would

$$[\Delta_i]_{ex}^{would}\varphi := \boxed{\scriptstyle S}((t(\Delta_i) \wedge \mathtt{e}_i) \to \varphi) \qquad (\mathtt{d}12)$$

Expected Could

$$[\Delta_i]_{ex}^{could}\varphi := [\Delta_i]_{ex}^{would}\varphi \wedge \Diamond_{\mathrm{S}}(t(\Delta_i) \wedge \mathtt{e}_i) \qquad (\mathtt{d}13)$$

Definitions (d12) and (d13) restrict the evaluation of ‘would’ and ‘could’ to those immediate future moments that the agent *expects* as continuations of the present. Adopting these modalities, we obtain two formal definitions of instrumentality corresponding to items (1) and (2) in Definition 4.11.

Potential φ-instrument for i (past and future)

$$[\Delta_i]_n^{p-instr^*}\varphi := \bigvee_{0\leq j<n} \blacklozenge^j(t(\Delta_i) \wedge \blacklozenge[\Delta_i]^{would}\varphi) \wedge [\Delta_i]_{ex}^{could}\varphi \tag{d14}$$

Excellent φ-instrument for i (past and future)

$$[\Delta_i]_n^{exc-instr^*}\varphi := [\Delta_i]_n^{p-instr^*}\varphi \wedge \bigwedge_{1\leq j\leq n} \blacksquare^j[\Delta_i]^{would}\varphi \tag{d15}$$

The conjectural element in (d14) and (d15) refers to the agent’s expectations *at the moment of evaluation.* By contrast, in evaluating the past, we must ignore the agent’s past expectations in selecting the agent’s relevant experience. In fact, a series of unexpected events in the past may have led the agent to the conviction that a particular action is a suitable instrument. Furthermore, the formalizations (d14) and (d15) differ from (d10) and (d11) via an additional conjunct expressing that the agent *expects* that she could guarantee φ by performing Δ (at the moment of evaluation). It can be straightforwardly observed that each excellent φ-instrument is also a potential φ-instruments and that (d11) and (d15) incorporate, respectively, (d10) and (d14) as their first conjunct.

Instrumentality—as discussed in this section—is a defeasible notion in three ways: first, depending on the length of the interval considered for evaluating the past, an instrument Δ may fail to qualify as a potential or an excellent φ-instrument once the interval is shortened and may fail to qualify as an excellent φ-instrument when the interval is extended. Second, with respect to the future, a φ-instrument may fail to remain a potential φ-instrument, either because an agent changes her expectations or, in the case of excellent instruments, because the instrument has failed to produce the desired end in the meantime. The third and foremost defeasible property of instrumentality arises through expectations: although an agent i expects that the (excellent) instrument will serve its intended end once again at the moment of evaluation, the *actual successor* is such that the instrument fails to deliver the purpose. Such cases reveal a discrepancy between i’s expectations and the actual future.

Figure 4.5 represents a model illustrating the three defeasibility aspects of instrumentality:

$$\begin{array}{llll}
\text{i)} & w_3 & \not\models & [\delta_i]_2^{exc-instr}p \rightarrow [\delta_i]_3^{exc-instr}p \\
\text{ii)} & w_4 & \not\models & [\delta_i]_2^{exc-instr}p \rightarrow \Diamond\kern-0.75em{\scriptstyle S}\kern0.3em[\delta_i]_3^{exc-instr}p \\
\text{iii)} & w_3 & \not\models & [\delta_i]_2^{ex-instr^*}p \rightarrow \boxed{\mathsf{A}}(t(\delta_i) \rightarrow p)
\end{array}$$

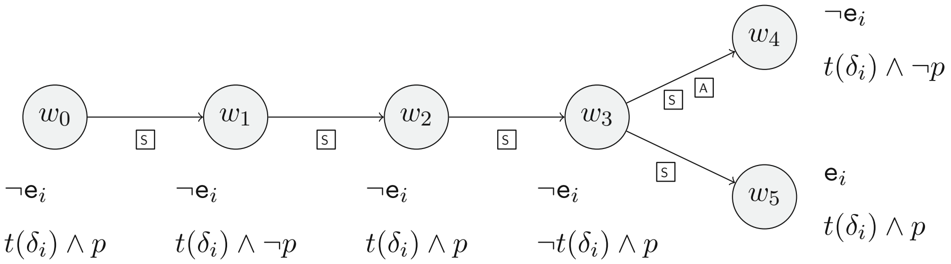


Figure 4.5: Three types of defeasibility concerning instrumentality relations.

Intuitively, (i) expresses that excellent instrumentality is not necessarily preserved when the past interval is extended; (ii) captures that excellent instrumentality is not necessarily preserved when time continues; and (iii) shows that an agent can be wrong about instrumentality due to actual future being unexpected.

4.7 Related Work and Future Research

Negation, PDL, and LAN. The logical behavior of *action negation* in complex actions has proven to be challenging. In this section, we briefly discuss some of these challenges in relation to LAN. Following the work by Broersen (2004), there are roughly two approaches to defining the complement '$-$' for action negation. These are referred to as the *universal* and the *relativized* approach.[18]

The universal approach defines action negation through a standard relational complement relative to the universal relation $W \times W$. In this approach, the complement of an action Δ is any potential transition between two moments except for those characterized by a performance of Δ, i.e., $\mathcal{R}_{\overline{\Delta}} := (W \times W) \setminus \mathcal{R}_{\Delta}$. Under this reading, the semantic interpretation of a formulae $[\overline{\delta}]\varphi$ is defined as:

$$\mathfrak{M}, w \models [\overline{\delta}]\varphi \text{ iff for all } v \in W, (w, v) \in (W \times W) \setminus \mathcal{R}_{\Delta},\ \mathfrak{M}, v \models \varphi$$

There is a problem with this approach. Under the universal approach, not performing a particular action at a particular moment may entail transitions between moments that are not reachable from the present moment (for instance, moments in the past). Figure 4.6 represents such a scenario: At moment w_0, the formula $[\overline{\delta}]\neg p$ is true even though i) it is impossible at w_0 to refrain from performing δ and ii) moment w_2 (for which $(w_0, w_2) \in \mathcal{R}_{\overline{\delta}}$) is unrelated and inaccessible from w_0.

Broersen (2004) adopts the relativized approach where negation is defined relative to those future moments *reachable* from the present moment of evaluation. One of the main features of action negation in the relativized approach is that actions such as

[18] We refer to the discussion by Bach (2010) for a critical assessment of negative action.

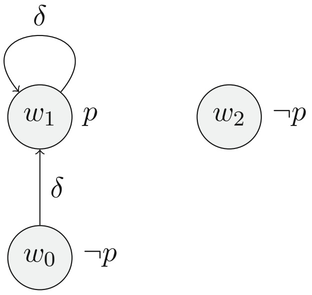


Figure 4.6: A PDL-like model $\mathfrak{M}$ illustrating the universal approach to action negation, where $\mathcal{R}_{\overline{\delta}}(w_0) = (W \times W) \setminus \mathcal{R}_\delta = \{(w_0, w_0), (w_0, w_2), (w_2, w_2)\}$ and $\mathfrak{M}, w_0 \models [\overline{\delta}]\neg p$.

"not opening the window" are taken as a "refraining from opening the window" which means that the agent in question is "actively ensuring that [the action] is not done" (2004, p.158). In other words, 'not opening the window' means that the agent *does anything but opening* the window at the given moment (cf. 'refraining' in STIT logics, Chapter 2.) Formally, relativized action negation takes the relational complement of all future moments that are reachable from the present moment through actions available to the agent at that moment. It is called "the reachable state-space" (Broersen, 2004). Since what is reachable from a given moment depends on the types of complex actions available in the language, relativized action negation receives different interpretations in different action languages. Actions of the language $\mathcal{L}^{\mathsf{Act}}$ contain only $\cup$ and $-$; thus, the reachable state-space consists of single transitions only. We adopt the PDL-like relational characterization of actions to the language $\mathcal{L}^{\mathsf{Act}}$.

Definition 4.12. *For each $\Delta \in \mathcal{L}^{\mathsf{Act}}$, the relation $\mathcal{R}^{pdl}_{\overline{\Delta}}$ is recursively defined by **S1**-**S3**:*

S1 *For each $\delta_i \in \mathsf{Act}$, $\mathcal{R}^{pdl}_{\delta_i} \subseteq W \times W$;*

S2 *For each $\Delta \in \mathcal{L}^{\mathsf{Act}}$, $\mathcal{R}^{pdl}_{\overline{\Delta}} = \bigcup_{\Gamma \in \mathcal{L}^{\mathsf{Act}}} \mathcal{R}^{pdl}_{\Gamma} \setminus \mathcal{R}^{pdl}_{\Delta}$;*

S3 *For each $\Delta, \Gamma \in \mathcal{L}^{\mathsf{Act}}$, $\mathcal{R}^{pdl}_{\Delta \cup \Gamma} = \mathcal{R}^{pdl}_{\Delta} \cup \mathcal{R}^{pdl}_{\Gamma}$.*

The semantic evaluation of the action modality $[\Delta]$ is defined accordingly:

- *$\mathfrak{M}, w \models [\Delta]\varphi$ iff for each $v \in \mathcal{R}^{pdl}_{\Delta}(w)$, $\mathfrak{M}, v \models \varphi$.*

As expressed by **S2**, the relativized action negation of Δ consists of all action transitions at w minus those transitions that correspond to a performance of action Δ at w.[19] We stress that the above definition expresses that the complement of an action Δ always

[19] Observe that there might be actions Γ that are identical to Δ at w, i.e., $\mathcal{R}^{pdl}_{\Gamma}(w) = \mathcal{R}^{pdl}_{\Delta}(w)$.

corresponds to some combination of action primitives. Namely, the reachable state-space is defined by transitions resulting from concrete primitive actions from which complex actions are built. Consequently, an agent's refraining from doing something must correspond to the agent doing some other concrete primitive action(s).

The semantic characterization given in Definition 4.12 is implied by our semantics. Theorem 4.4 demonstrates this. The following lemma shows that the relation $\mathcal{R}_\Delta := \mathcal{R}_{\boxed{\mathsf{S}}} \cap (W \times W_{t(\Delta_i)})$ characterizes the PDL-like *defined* modal formula $[\Delta]\varphi \in \mathcal{L}^{\mathsf{LAN}}$.

Lemma 4.8. *Let $\mathfrak{M}$ be an arbitrary* LAN*-model and let for each $\Delta \in \mathcal{L}^{\mathsf{Act}}$, $\mathcal{R}_\Delta := \mathcal{R}_{\boxed{\mathsf{S}}} \cap (W \times W_{t(\Delta)})$. Then, $\mathfrak{M}, w \models [\Delta]\varphi$ iff for each $v \in \mathcal{R}_\Delta(w)$, $\mathfrak{M}, v \models \varphi$.*

Proof. We recall that $\mathcal{R}_\Delta(w) = \{v \in W \mid (w, v) \in \mathcal{R}_\Delta\}$. It suffices to observe the following equivalences: for each $v \in \mathcal{R}_\Delta(w)$, $\mathfrak{M}, v \models \varphi$ iff for each $v \in \mathcal{R}_{\boxed{\mathsf{S}}}(w) \cap W_{t(\Delta)}$, $\mathfrak{M}, v \models \varphi$ iff for each $v \in \mathcal{R}_{\boxed{\mathsf{S}}}(w)$, if $v \in W_{t(\Delta)}$, then $\mathfrak{M}, v \models \varphi$ iff for each $v \in \mathcal{R}_{\boxed{\mathsf{S}}}(w)$, if $\mathfrak{M}, v \models t(\Delta)$, then $\mathfrak{M}, v \models \varphi$ (Lemma 4.1) iff for each $v \in \mathcal{R}_{\boxed{\mathsf{S}}}(w)$, if $\mathfrak{M}, v \models t(\Delta) \to \varphi$ iff $\mathfrak{M}, w \models \boxed{\mathsf{S}}(t(\Delta) \to \varphi)$ iff $\mathfrak{M}, w \models [\Delta]\varphi$. QED

Theorem 4.4. *Let $\mathfrak{F} = \langle W, \mathcal{R}_{\boxed{\mathsf{S}}}, \mathcal{R}_{\boxed{\mathsf{A}}}, \{W_{\mathsf{d}_i^\delta} \mid \mathsf{d}_i^\delta \in \mathcal{L}^{\mathsf{LAN}}\}, \{W_{\mathsf{v}_i} \mid i \in \mathsf{Agents}\}\rangle$ be a* LAN*-frame and let $W_{t(\Delta)}$ be as in Definition 4.5. For each $\Delta \in \mathcal{L}^{\mathsf{Act}}$, let $\mathcal{R}_\Delta = \mathcal{R}_{\boxed{\mathsf{S}}} \cap (W \times W_{t(\Delta)})$. Then, $\mathcal{R}_\Delta$ satisfies properties **S1**-**S3**.*

Proof. First, we observe that for each $w \in W$:

$$(\dagger) \qquad \bigcup_{\Delta \in \mathcal{L}^{\mathsf{Act}}} \mathcal{R}_\Delta(w) = \mathcal{R}_{\boxed{\mathsf{S}}}(w)$$

Left-to-right. Trivial by definition of $\mathcal{R}_\Delta$. Right-to-left. Assume $v \in \mathcal{R}_{\boxed{\mathsf{S}}}(w)$ for some $v \in W$ and suppose towards a contradiction that $v \notin \bigcup_{\Delta \in \mathcal{L}^{\mathsf{Act}}} \mathcal{R}_\Delta(w)$. Hence, there is no $\Delta \in \mathcal{L}^{\mathsf{Act}}$ such that $v \in \mathcal{R}_\Delta(w)$. We know $v \in W$ by assumption. Furthermore, we know that for each atomic $\delta \in \mathsf{Act}$, $W = W_{t(\delta_i)} \cup W_{t(\overline{\delta_i})}$. Take an arbitrary $\delta_i \in \mathsf{Act}$. Assume $v \in W_{t(\delta_i)}$. Then, since $(w, v) \in \mathcal{R}_{\boxed{\mathsf{S}}}$ we know $v \in \mathcal{R}_{\delta_i}(w) = \mathcal{R}_{\boxed{\mathsf{S}}}(w) \cap W_{t(\delta_i)}$. Contradiction. Assume $v \in W_{t(\overline{\delta_i})}$. Then, by the same reasoning we know $v \in \mathcal{R}_{\overline{\delta_i}}(w) = \mathcal{R}_{\boxed{\mathsf{S}}}(w) \cap W_{t(\overline{\delta_i})}$. Contradiction.

We use the fact (†) in proving that $\mathfrak{F}$ satisfies **S3**.

S1 Since $W_{t(\delta_i)} \subseteq W$ and $\mathcal{R}_{\boxed{\mathsf{S}}} \subseteq W \times W$ we have $\mathcal{R}_{\delta_i} = \mathcal{R}_{\boxed{\mathsf{S}}} \cap (W \times W_{t(\delta_i)}) \subseteq W \times W$.

S2 $\mathcal{R}_{\Delta \cup \Gamma} = \mathcal{R}_{\boxed{\mathsf{S}}} \cap (W \times W_{t(\Delta \cup \Gamma)}) = \mathcal{R}_{\boxed{\mathsf{S}}} \cap (W \times (W_{t(\Delta)} \cup W_{t(\Gamma)})) = \mathcal{R}_{\boxed{\mathsf{S}}} \cap (W \times W_{t(\Delta)}) \cup \mathcal{R}_{\boxed{\mathsf{S}}} \cap (W \times W_{t(\Gamma)}) = \mathcal{R}_\Delta \cup \mathcal{R}_\Gamma$.

S3 $\mathcal{R}_{\overline{\Delta}} = \mathcal{R}_{\boxed{\mathsf{S}}} \cap (W \times W_{t(\overline{\Delta})}) = \mathcal{R}_{\boxed{\mathsf{S}}} \cap (W \times W \setminus W_{t(\Delta)}) = \mathcal{R}_{\boxed{\mathsf{S}}} \setminus (\mathcal{R}_{\boxed{\mathsf{S}}} \cap (W \times W_{t(\Delta)})) = \mathcal{R}_{\boxed{\mathsf{S}}} \setminus \mathcal{R}_{t(\Delta)} = \bigcup_{\Gamma \in \mathcal{L}^{\mathsf{Act}}} \mathcal{R}_\Gamma \setminus \mathcal{R}_{t(\Delta)}$, by (†). QED

Last, there is a set of formulae that is satisfiable on a LAN frame, which is not satisfiable on any model of relativized action negation $\mathfrak{M}^{pdl}$ (where the modal action formulae are semantically characterized by **S1-S3**). Consider the following set of formulae provided by Broersen (2003, p.82):

$$\Sigma = \{\langle\overline{\delta^1}\rangle\neg p\} \cup \{[\delta^i]p \mid \delta^i \in \mathsf{Act}\}$$

where $\{\delta^1, \delta^2, \delta^3, \dots\} = \mathsf{Act}$. The set Σ is not satisfiable on any $\mathfrak{M}^{pdl}$ although each finite subset $\Sigma' \subseteq \Sigma$ is. In other words, the corresponding PDL logic is not compact and a fortiori, not strongly complete. The reason is that the complement of δ^1 must correspond to some positive atomic actions $\delta^i, \dots, \delta^j \in \mathsf{Act}$ but the whole set Σ expresses that all primitive actions $\delta^i \in \mathsf{Act}$ lead to p. Any finite subset Σ' is satisfiable since either $\langle\overline{\delta^1}\rangle\neg p$ is left out in $\Sigma' \subseteq \Sigma$ and thus all primitive actions lead to p, or $[\delta^i]p \in \Sigma$ is left out in Σ' such that we may provide a model such that $\delta^i = \overline{\delta^1}$ leads to $\neg p$.

The logic LAN does satisfy the set Σ. As proven in Section 3.3, LAN is strongly complete (and thus compact). The reason is simple: our logical framework allows for transitions (i.e., reachable moments) that do not correspond to any atomic action. By definition, such a moment corresponds to negated actions. Thus, we allow for transitions to moments where no atomic action witnesses are satisfied. In satisfying Σ, we can define such a moment accessible from the moment of evaluation to satisfy $\langle\overline{\delta^1}\rangle\neg p$. One can think of this non-action transition as the transition in which the agent completely refrains from acting, e.g., where nature causes the transition. In fact, this is where our approach differs from dynamic action logics adopting a relativized action negation and STIT logics. There, every change corresponds to an agentive change: when an agent remains passive, being passive corresponds to an action, respectively, choice. Consequently, in those approaches, change is agentively exhaustive. This differs from the approach by von Wright (1963a) and the logic introduced in this chapter. In LAN, 'being passive' means that the agent does not actively interfere with the course of nature. As Hilpinen (1997) puts it: "If an action is regarded as an interference with a natural course of events (a course of events unaffected by agents) in the way proposed by von Wright, then the concept of passivity or 'zero action' is meaningful. [...] Actions are regarded here as analogous to tools which the agent can use to work on a situation" (p.20). As shown in this section, our approach has some interesting technical consequences: LAN is a variant of dynamic deontic logic that includes 'passivity', consequently rendering the logic strongly complete.

Other Action Logics. Belief-Desire-Intention (BDI) systems are rooted in the philosophy of practical reasoning (Bratman, 1981) and deal with deliberation and planning (Rao and Georgeff, 1995). BDI logics are multi-modal multi-agent logics whose language contains modalities for beliefs, desires, and intentions, as well as temporal modalities (Meyer et al., 2015; Rao and Georgeff, 1998). In order to capture different types of agents, the logics can be extended with properties expressing various logical relations between belief, desire, and intention. For instance, the property 'realism' signifies that if an agent believes φ, she also has an intention towards φ. We refer to the work of Meyer et al.

(2015) for a discussion of BDI logic in relation to STIT logic. As indicated in Chapter 1, this thesis does not deal with instrumentality in the context of planning and BDI logics. Instead, we investigated instrumentality relations as subject to norms in this chapter. The study of BDI logics and norms of instrumentality is a promising future research direction.

Hughes et al. (2007) formally investigate means-end statements in the context of practical reasoning. They provide a formal semantics of such statements using Propositional Dynamic Logic (without a sound and complete proof system). Their action language consists of disjunctive and sequential action, extended with the test operator. They formalize sufficient and necessary means. The main difference with our approach is the involvement of action-negation in LAN, which allows us to express deontic modalities in terms of violations.

The formalism developed by Åqvist (2002) employs a rich formal language containing seven modal operators. Åqvist shows how to approximate other formal accounts of action, e.g., accounts by Belnap, Von Kutschera, von Wright, and Segerberg. The main difference with our approach is that we employ a minimal modal language containing action and violation constants, which suffices for defining agency modalities and deontic modalities. An analysis of whether our work embeds or can be embedded in the formalism of Åqvist (2002) remains to be determined.

Segerberg (2002) also provides a (non-deontic) action logic in the tradition of von Wright's theory of action, time, and change. The language contains temporal operators and only one action operation: sequential action. Furthermore, Segerberg differentiates between the agentive "bringing it about that" and the non-agentive "coming about that". Conceptually, the distinction between agentive and non-agentive change is promising for investigating responsibility and moral culpability. This line of research is not further pursued in this thesis.

Åqvist (2002) proposes a list of criteria that any action logic should address. We briefly recall each and discuss whether LAN satisfies these criteria; cf. (Åqvist, 1974). First, the logic must adopt von Wright's ideas that action involves change, transformations of states, or transitions. This theory lies at the heart of LAN (Section 4.1). Second, the language must be able to differentiate between individual acts and generic acts. The language of LAN provides this distinction by having a general algebra of actions $\mathcal{L}^{\mathsf{Act}}$, which is mapped to action-tokens in $\mathcal{L}^{\mathsf{LAN}}$ witnessing the performance of concrete actions by agents. Third, the philosophy of action maintains a strict distinction between performing and omitting, and action logic must be able to differentiate the two. In Section 4.2, we showed how von Wright's four elementary action types (as well as forbearance) can be modeled in LAN. The logic must capture the notion of "bringing about something". This topic was not addressed in this chapter. However, we point out that the notions of 'would', 'could', and 'will' capture forms of agency that relate to the concept of "bringing about that". Last, the logic must be empirically testable, i.e., it must have an application. We did not address this last criterion. In the next chapter, we develop a logic (based on

LAN) and apply it to an analysis of a deontic theory developed by a prominent South Asian Sanskrit philosopher.

Combining STIT and PDL-like Logics. Various authors have proposed systems that combine the STIT formalism with PDL-like approaches to action logic. Most notably, Segerberg (2002) proposes a dynamic action logic that captures both von Wright's conception of elementary action (cf. Section 4.1) and various agentive notions related to STIT. Others, such as Broersen (2014) and Xu (2010), adopt STIT as the basic formalism, extending it with explicit actions or event types. It goes beyond the objectives of Part I and II to investigate whether LAN can be integrated with the STIT formalism.

Deontic Action Logics and Green States Giordani and Canavotto (2016) introduce the logic ADL, which contains more syntactic diversity than most standard dynamic deontic logics. For instance, they distinguish what is ideal in general from what is ideal in a specific situation. Furthermore, they distinguish between an action ("opening the door") and the result of an action ("the door is open"). The latter does not necessarily presuppose the former: that the door is open may or may not be the result of opening the door. In contrast, LAN adopts a relatively simple language in which actions are reduced to constants that witnesses a successful performance of that action. The logic ADL belongs to the relativized action negation tradition (see page 161). Giordani and Pascucci (2022) propose a deontic action logic similar to LAN and the approach by (Giordani and Canavotto, 2016). They adopt the idea of an action-witness by letting '$done(\delta)$' be a state of affairs, denoting that the performance of action δ is done.

The deontic action logic $\mathcal{DAL}$, developed by Trypuz and Kulicki (2015), likewise adopts an algebra of actions. In contrast to PDL-like logics such as the one by Meyer (1988), it does not employ action modalities. However, it incorporates complex actions directly into the deontic operators for forbidden and permitted actions. For instance, the formula $\mathcal{P}(\delta\&\overline{\gamma})$ expresses that "performance of the complex action $\delta\&\overline{\gamma}$ is permitted". Furthermore, Trypuz and Kulicki (2015) also investigate the concept of performing no action at all, i.e., the zero action (see our discussion on page 164). In $\mathcal{DAL}$, they impose the constraint that not acting at all cannot be void of deontic value. That is, it is always either permitted or forbidden to not act at all.

Craven and Sergot (2008) propose a deontic extension of the action description language C^+ for defining labeled transition systems where transitions and states can be labeled 'permitted'. In the deontic extension of C^+, sets of permitted states are defined separately from the sets of permitted transitions. These permitted states, respectively, transitions are then called "green" (in contrast to the non-compliant "red" states and transitions). The additional expressivity of coloring states and transitions separately allows Craven and Sergot (2008) to formulate deontic constraints such as the "green-green-green constraint," which expresses that "a green (permitted, acceptable, legal) transition in a green (permitted, acceptable, legal) state always leads to a green (acceptable, legal, permitted) state" (p.228). Furthermore, coloring states and transitions enables the representation of system norms and agent-specific norms. The former "regulate the interactions of multiple, independently acting agents in a multi-agent computer system" whereas the latter takes into account "what an agent can actually sense or perceive and the actions that it can actually perform" (Sergot, 2008, p.16). We refer to Craven and Sergot (2008) for a discussion of the essential differences between the deontic extension of C^+ and Deontic Dynamic Logics, such as the one developed by Meyer (1988). Here, we point out that LAN only allows for identifying permitted states of affairs and actions by reference to states witnessing violation constants. It is left to future work to investigate whether labeling states (outcomes) and transitions (actions) separately would lead to a more refined analysis of norms instrumentality in the context of LAN. To the best of our knowledge, an investigation of norms of instrumentality has yet to be conducted in the deontic extension of C^+.

Lomuscio and Sergot (2003) adopt a modal logic approach to model correct and compliant behavior of agents with reference to "green states". An essential difference between the work by Lomuscio and Sergot (2003) and LAN (and the deontic extension of C^+ discussed above) is that no explicit action language is involved in the former. Like LAN, the logic developed by Lomuscio and Sergot (2003) allows for an Andersonian reduction in which obligations are defined in terms of a "green" constant together with a $\Box$ modality.

The logic LAN belongs to the Deontic Dynamic Logic tradition. The main difference between our approach and existing approaches to dynamic logics is that we use action constants to define an "Andersonian" reduction of *action modalities*. The upshot of our approach is that we only need one modality $\boxed{\scriptstyle S}$ together with action and violation constants to define a large class of agentive and deontic concepts (recall that the actual successor modality $\boxed{\scriptstyle A}$ is primarily needed for defining the agentive concept of "will"). It is left to future work to determine to what extent other agentive concepts, such as those proposed by Åqvist (2002) and Segerberg (2002), can be expressed in LAN.

Sequential action in a deontic context. In the language of LAN, we can define sequences of actions, denoted by the operation ';', as

$$[\Delta_i;\Gamma_i]\varphi := \Box(t(\Delta_i) \to [\Gamma_i]\varphi)$$

provided Δ_i does not contain any sequential action. The reason for this side condition is that the translation function $t(\cdot)$ is presently only defined for actions composed of primitive actions and the operations $\cup$ and $-$. Extending the language $\mathcal{L}^{\mathsf{Act}}$ and translation function $t(\cdot)$ to include the sequence operator ; is not trivial. The reason is that we need to define the interaction between action negation and action sequence. This becomes clearer in a deontic setting where commands are defined in terms of actions and violations. An obligation to perform the sequence $\Delta_i;\Gamma_i$ means that $\overline{\Delta_i;\Gamma_i}$ leads to a violation, i.e., $[\overline{\Delta_i;\Gamma_i}]\mathsf{v}_i$. There are several ways to define this complement. For instance, we may interpret $[\overline{\Delta_i;\Gamma_i}]\mathsf{v}_i$ as "any reachable moment which is not a moment reached by agent i first performing Δ and subsequently performing Γ is a state of violation". This interpretation is arguably too strong since it resists CTD reasoning, i.e., after an obligation is violated, every future state is a violation state, and no (CTD) obligations are possible. Alternatively, we may interpret $\mathcal{O}_i[\Delta; \Gamma]$ as "a violation occurs when agent i performs $\overline{\Delta}$ (making it impossible to finish the sequence correctly) and when i started with Δ but subsequently performs $\overline{\Gamma}$". In other words, a violation occurs at the first moment when the obligation cannot be fulfilled anymore. This is the approach proposed by Meyer (1988). Given this reading, the agent may find herself in a violation state and still have a CTD obligation. We refer to the work of Anglberger (2008) for a critical assessment of the dynamic deontic logic developed by Meyer (1988). A thorough analysis of sequential action in the context of deontic modalities is not pursued in this thesis.

Open question 4.1. *Investigate the notion of sequential action in deontic contexts using the action-reductionist approach of* LAN*.*

* * *

In this chapter, we addressed the question of instrumentality, or means-end, statements in the context of normative reasoning. We provided a Logic of Action and Norms called LAN to reason about such statements. We identified a norm category called *norms of instrumentality*, formalized it in the language of LAN (Objective 1), and investigated the logical relations between this norm category and the well-known categories of *norms to be* and *norms to do* (Objective 2). Furthermore, based on the work of von Wright, we discussed possible extensions of LAN in which various kinds of more refined instrumentality statements can be formalized. Last, we argued that LAN subsumes the action relativized action negation approach from PDL.

CHAPTER 5

An Application to Sanskrit Philosophy

The logical analysis of normative reasoning is a relatively young field of research (Hilpinen and McNamara, 2013). By contrast, the school of Mīmāṃsā—one of the most important schools of Indian philosophy—has a long and rich history of investigating normative reasoning. The school was active and influential for over two millennia, shaping many related areas in the Sanskrit cosmopolis.[1] It focuses on the exegesis and systematization of the *prescriptive parts of the Vedas*, the sacred texts of what is now called Hinduism. The Mīmāṃsā consider the Vedas to be without any human or divine author, and, what is more, they assume all Vedic commands to be jointly consistent. Consequently, Mīmāṃsā authors invested much intellectual effort in rationally interpreting Vedic commands, explaining what must be done in the presence of *seeming* conflicts. The result is a vast body of rigorously structured theories of normative reasoning based on general principles of inference. Due to their highly systematic nature, these principles have been applied in many other fields and are still used in Indian jurisprudence (McCrea, 2010). Thus, Mīmāṃsā reasoning lends itself naturally to logical analysis.

Despite their undeniable importance, most Mīmāṃsā doctrines are still unexplored or misunderstood. In this chapter, we use symbolic logic to formalize and obtain a deeper understanding of the deontic theory developed by one of the Mīmāṃsā's most central authors: *Maṇḍana* miśra(ca. 8th c. CE, henceforth, Maṇḍana). Our first objective is formulated accordingly:

Objective 1. *Provide an adequate formal logic modeling Maṇḍana's deontic theory.*

[1] Mīmāṃsā doctrines have influenced Sanskrit philosophy, theology, and law more than any other system of thought. See the work by McCrea (2010) and McCrea (2008) on Mīmāṃsā influence on law, aesthetics, respectively theology.

In answer to this objective, we develop the Logic of Maṇḍana, LM for short.

Maṇḍana's deontic theory is unique in the Mīmāṃsā tradition because it contains a *deontic reduction*: i.e., a uniform reduction of all Vedic commands to purely descriptive statements about desires, outcomes, and instruments. For instance, the Vedic command "If one desires rain, one should perform the Kārīri ritual" is reduced to the descriptive statement "the Kārīri is an instrument for attaining rain". A central feature of this reduction is that different commands are reduced to the singular notion of *instrument*. An immediate question is whether the validity of relevant normative reasoning principles developed by the Mīmāṃsā—called *nyāyas*—is preserved through Maṇḍana's reduction.

Objective 2. *Employ the logic* LM *to enhance our understanding of Maṇḍana's deontic reduction and its relation to general Mīmāṃsā principles.*

In line with the above objective, we apply LM to model Maṇḍana's solution to the Śyena controversy. The Śyena is a ritual in which the so-called Soma beverage is offered. Its putative result is the death of the sacrificer's enemy. The controversy is due to the fact that the Vedas appear to both prescribe the Śyena and prohibit the infliction of harm on any living being, thus yielding a conflict. Finding a solution to the controversy proved challenging for many Mīmāṃsā scholars. Furthermore, the Śyena controversy can be seen as a millennia-old counterpart to the deontic paradoxes that drive developments in modern deontic logic (Hilpinen and McNamara, 2013) (see Chapter 1).

Objective 3. *Employ* LM *to model Maṇḍana's solution to the Śyena controversy.*

This chapter is part of a series of logics developed for various Mīmāṃsā authors (Ciabattoni et al., 2015; Freschi et al., 2017; van Berkel et al., 2019; Freschi et al., 2019; Lellmann et al., 2021; van Berkel et al., 2021a; Freschi and Pascucci, 2021; van Berkel et al., 2022a). We believe that the Mīmāṃsā school can offer new stimuli for the deontic logic community, challenging commonly accepted design choices, such as the interdefinability of obligations and prohibitions and the presence and absence of certain deontic principles. To determine its legitimacy in the contemporary field of deontic logic, we must determine whether the logic LM can deal with the benchmark challenges posted by the community (Hilpinen and McNamara, 2013).

Objective 4. *Evaluate* LM *on a set of deontic paradoxes from the deontic logic community.*

We will see that the logic LM consistently addresses these deontic paradoxes when reformulated in terms of Maṇḍana's *reduction*. The solution strategy of Maṇḍana's theory resembles a well-known strategy in modern deontic logic, i.e., adopting a logic of actions (Meyer, 1988) and reducing commands to statements concerning rewards and sanctions (Anderson and Moore, 1957). These encouraging results may be due to the depth of the deontic theory that underlies the formalized logic. Namely, LM is grounded in a fully developed philosophical and juridical system of thought.

Contributions. In this chapter, we address the above four objectives. The majority of the present chapter consists of results first published in (van Berkel et al., 2021a; van Berkel et al., 2022a). The former work introduces the initial logic modeling Maṇḍana's deontic theory. In (van Berkel et al., 2022a), the logic was modified due to the discovery of an additional Mīmāṃsā principle (cf. **P4** on page 190). The resulting logic LM was shown sound and strongly complete (Objective 1). It was shown that Maṇḍana's deontic theory can be accurately formalized in LM, and a formal investigation of Maṇḍana's theory in relation to a class of Mīmāṃsā principles was provided (Objective 2). Furthermore, Maṇḍana's solution to the Śyena controversy was consistently formalized and discussed (Objective 3). Last, in (van Berkel et al., 2022a) LM was employed to formally analyze several deontic paradoxes from the deontic logic literature (Objective 4).

Differences. The following parts of the chapter are novel: Section 5.1 contains a discussion of action and agency in Mīmāṃsā. In Section 5.3, the full proofs of soundness and completeness of the logic LM are presented. These proofs were only sketched in the previous works. In Section 5.4, the formalization of Mīmāṃsā principle **P2** in LM has been changed to range over actions instead of results, thus changing the formal analysis accordingly. Furthermore, an analysis of Jørgensen's dilemma in the context of Maṇḍana's deontic theory has been provided in Section 5.6. Last, a more extensive discussion of related work is provided in Section 5.7.

Outline. The chapter is organized as follows: Section 5.1 contains an introduction to Mīmāṃsā in general and to Maṇḍana's deontic theory in particular. In Section 5.2, we provide the modal logic LM tailored to Maṇḍana's doctrine. The logic LM is shown sound and complete in Section 5.3. After that, in Section 5.4, we discuss which Mīmāṃsā properties are valid in the context of LM. In Section 5.5, we put the logic to work and formalize Maṇḍana's solution to the Śyena controversy. In Section 5.6, we evaluate LM on a set of deontic paradoxes. Last, we discuss related work Section 5.7.

5.1 An Introduction to Mīmāṃsā and Maṇḍana

This section serves as the theoretical foundation of our formal analysis of Maṇḍana's reduction of normative reasoning to instrumentality statements. We first provide a brief introduction to Mīmāṃsā and situate Maṇḍana in this historical context. We then provide an account of Maṇḍana's deontic theory.

5.1.1 Mīmāṃsā

Mīmāṃsā is one of the main schools of Sanskrit philosophy. It is the only one focusing on the analysis of norms. Thriving for over two millennia—from the last centuries BCE to the 20th century—Mīmāṃsā focuses on the exegesis of the prescriptive portions of the Vedic sacred texts. A commonly used example of a Vedic prescription is "the one who is desirous of heaven should sacrifice with the New and Full Moon sacrifices" (Freschi, 2010, p.421). Mīmāṃsā authors devised a system of rules called *nyāyas*, which are meant to apply to any normative text. The *nyāyas* are used to understand the Vedas independently of any (super)human authority or mediation.[2] Mīmāṃsā authors agree that the Vedas are a consistent corpus of rules, which means that what might look like a conflict can be consistently resolved by applying the correct *nyāyas*.

Different Mīmāṃsā authors adopt different views, interpreting Vedic commands in different ways. Still, they all recognize the authority of the following two works: Jaimini's *Mīmāṃsā Sūtra* (or *Pūrva Mīmāṃsā Sūtra*, henceforth PMS, approximately 250 BCE) and Śabara's *Bhāṣya* commentary thereon (henceforth ŚBh, approximately 5th c. CE). We refer to this shared foundation as "common Mīmāṃsā".[3] Of particular importance are the following three authors, who originated different subschools in Mīmāṃsā:

- Kumārila (ca. 7 CE): considered to be the founder of the Bhāṭṭa subschool;
- Prabhākara (ca. 7 CE): a younger contemporary of Kumārila, considered to be the founder of the Prābhākara subschool;
- Maṇḍana (8 CE): authored independent treatises on various issues (especially on the nature of prescriptions) and innovated the Bhāṭṭa school.

Common Mīmāṃsā classifies the commands encountered in the Vedas into *prescriptions* and *prohibitions*. In general, such commands are directed at human beings. In what follows, we adopt the more general term agents. A command always contains an action. Prescriptions are often about sacrifices and are further differentiated based on the type of duty enjoined: *nitya-karman* 'fixed sacrifices' are to be performed every single day; *naimittika-karman* 'occasional sacrifices' are to be performed only on given occasions (e.g., a sacrifice to be performed on the birth of a child); *kāmya-karman* 'elective sacrifices' are to be performed solely if one wishes to obtain their result. These prescriptions have *varying deontic strength*: an agent may not omit the performance of fixed and occasional sacrifices (various authors provide different reasons for this), whereas the performance of elective sacrifices can be omitted without any adverse consequence (apart from not getting the intended result). For some authors, such as Prabhākara, all prescriptions are obligations, whereas others, such as Kumārila, further divide prescriptions into obligations and elective duties. We come back to this in Section 5.4. Furthermore, prescriptions are understood in relation to eligibility conditions (*adhikāra*). These include the agent's belonging to a particular class of living beings, the agent's ability to perform the prescribed action, and the agent's desire for the action's intended result.

[2] In this sense, Mīmāṃsā authors differ from other thinkers offering systematic interpretations of sacred texts. For instance, Talmudic normative reasoning depends on the mediation of a rabbi who applies the Talmud (Abraham et al., 2011), whereas Mīmāṃsā reasoning depends on abstract principles.

[3] The PMS is divided into books, chapters and aphorisms. We adopt the referencing style common in Sanskrit philosophy scholarship by indicating the number of the book, chapter, and section, respectively. For example, PMS 1.1.1 indicates the first aphorism of the first chapter of the first book.

Prohibitions form a separate category of commands, and Mīmāṃsā authors distinguish between prohibitions 'regarding the person' (*puruṣārtha*), i.e., applying to the agent throughout the agent's life, and those 'regarding the sacrifice' (*kratvartha*), i.e., applying only to the specific situation of the sacrifice. An analogy would be the command "do not kill", which applies to an agent's entire life, and the command "do not dress informally", which applies only in specific settings.

Obeying a prescription generates a positive result, namely, the result of the prescribed action. Prescriptions presuppose one's desire for this result. When an explicit desire or result is absent, a standard desire for happiness is postulated. Violating a prescription implies the absence of these results. Conversely, the observance of a prohibition generates no result, whereas a violation leads to a sanction, typically the accumulation of bad karma. Thus, prohibitions cannot be defined in terms of prescriptions or obligations—i.e., as a negative obligation—because the observance and transgression of these two types of commands have different consequences.[4] For instance, suppose that there are two simultaneous commands: a prohibition to lie and a negative obligation not to tell lies. Although the effect of compliance may seem the same for these commands, in the case of the negative obligation an additional mental act (*mānasakarman*) is involved: the resolve to *not* lie. It is this mental act that leads to a result, e.g., the accumulation of happiness. The difference between negative obligations and prohibitions is extensively discussed by Mīmāṃsā authors.

Prabhākara and Kumārila. Maṇḍana's deontic theory is strikingly different from those Mīmāṃsā authors that come before him. In order to make this clear, we briefly discuss the systems of Prabhākara and Kumārila.

Prabhākara's system is eminently deontic: agents follow commands because they are enjoined. Such agents recognize that they are enjoined because of the eligibility conditions in Vedic commands. For instance, the command "one who desires cattle should sacrifice with the Citrā" identifies the one who desires cattle as the enjoined agent to which the duty to sacrifice with the Citrā applies. In Prabhākara's system, once the eligibility conditions are met, the sacrifice *must* be performed. This means that, unlike in common Mīmāṃsā, for Prabhākara there is no normative distinction between fixed/occasional and elective sacrifices: they are all obligations.

Kumārila's deontic theory differs from Prabhākara's on the interpretation of elective sacrifices (*kāmya-karman*). Kumārila interprets these prescriptions as not properly binding as the agent in question can refrain from performing the enjoined action without any normative consequences. For Kumārila, elective duties only give a guaranteed way to bring about the desired result. They are, so to say, Vedic means-end recipes for obtaining certain desired results. Thus, an agent can ignore the desire for a specific result of an elective ritual but not those of fixed and occasional rituals. This is because the latter lead to happiness, which is, according to Kumārila an aspiration characterizing every human being.

[4] This is contrary to common deontic logic practice (Hilpinen and McNamara, 2013).

Actions and Mīmāṃsā. Following Freschi (2010), the Mīmāṃsā take action foremost as effort, i.e., the initiation of activity. Śabara (cf. common Mīmāṃsā) uses the noun *bhāvanā* to denote 'undertaking an activity in general'. For him, *bhāvanā* is the general activity of bringing about a certain aim. In fact, every action has an aim for the Mīmāṃsā. In the cases of Vedic prescriptions, this "bringing about" takes place through performing the prescribed sacrifice. Maṇḍana also adopts the view that all action is accompanied by an active "bringing about that" (*bhāvanā*) together with effort. The latter consists of a volitional act that initiates an act (e.g., a will determination). Furthermore, effort itself is influenced by desire and aversion. In fact, for Maṇḍana, any action serves the purpose (aim) of achieving pleasure or avoiding pain. Thus, the three main components are *desire*, *purpose*, and *action*. We may take the purpose that the action serves as that which is desired. To illustrate, reconsider the prescription "the one who desires cattle should sacrifice with the Citrā". Roughly, the object of desire is cattle, and the action is the Citrā ritual, which serves the purpose of bringing about the attainment of cattle. Some aims, such as heaven, are not directly obtained after the action concludes. Heaven is only acquired after one's death. In order to account for actions that bring about results in a distant future, the Mīmāṃsā postulated *apūrva* which is the immediate outcome of the performed action eventually leading to the action's aim (such as heaven). See (Freschi, 2010) for a more detailed introduction to the analysis of agency in the Mīmāṃsā school.

5.1.2 Maṇḍana

Maṇḍana's account of normative reasoning breaks with the Mīmāṃsā tradition. According to Maṇḍana, fixed and occasional duties, elective duties, and prohibitions can be expressed solely in terms of desires, outcomes, and instruments. Maṇḍana's approach is, thus, a *deontic reduction*: a reduction of all Vedic commands to purely descriptive statements of instrumentality. To illustrate this, consider the prescriptive statement "one who desires to kill their enemy should perform the Śyena sacrifice". On Maṇḍana's account, this command is reduced to the descriptive statement "the Śyena is an instrument for killing one's enemy". One of the central features of the reduction is that different commands are reduced to the singular notion of *instrument*.[5] An instrumentality relation signifies a relation between an action and a result: the action is regarded as the instrument leading to the intended result (see Chapter 4 for an analysis of instrumentality). The result is a state of affairs. Hence, in contrast to Prabhākara and Kumārila, for Maṇḍana, deontic concepts such as obligations and prohibitions arise from differences in instrumentality relations.

The uniform language employed in the reduction may suggest that different commands are reduced to instrumentality statements with the same normative status. However, to maintain the desired distinction between fixed/occasional duties, elective duties, and prohibitions, Maṇḍana adopts two constraints involving the accumulation and the reduction of bad karma (*pāpa* in Sanskrit). First, fixed and occasional duties describe

[5] In Section 5.5, we discuss how Maṇḍana deals with controversial commands like the one about Śyena.

actions instrumental to the reduction of bad karma. To distinguish those duties from other types of instruments that fulfill desires, Maṇḍana argues at length that the desire for the *reduction* of bad karma is a unique desire shared by every rational being (cf. Kumārila's postulate that happiness is universally desired). Second, to ensure that prohibitions retain their prohibitive strength, Maṇḍana argues that prohibited actions lead to strongly undesirable outcomes whose undesirability is incommensurably greater than any desirable result, including the desire to reduce bad karma. For Maṇḍana, this universally undesirable result is the *accumulation* of bad karma. Elective duties are, then, taken to describe instrumentality relations between actions and results for those actions that neither lead to the reduction nor to the accumulation of bad karma directly. These desires are called worldly desires, such as the desire for more cattle.

Since obligations and elective duties lead to something desirable, they are grouped under the term *iṣṭasādhana*, i.e., "instrument to something desirable" (with the reduction of bad karma being universally desirable). Prohibitions are actions instrumental to something strictly undesirable and are called *aniṣṭasādhana*, i.e., "instruments to something undesirable" (with 'an-' being the Sanskrit equivalent to the English prefix 'un-').

Maṇḍana does not claim that bad karma is something that *ought to be* reduced or avoided. Instead, he argues that, in a conflict between worldly and karma desires, no rational being would prefer the former over the latter (we come back to this when discussing the Śyena controversy in Section 5.5). Thus, Maṇḍana proposes a unifying theory for normative reasoning that reduces all command types to instrumentality statements about actions leading to results. We call it Maṇḍana's deontic reduction.

5.1.3 Our Methodology

The logic LM results from an interdisciplinary collaboration between scholars of logic, computer science, and Sanskrit philosophy. It is worth taking a closer look at the methodology employed.

The aim is to represent the reasoning of Maṇḍana faithfully. This means we want to impose as few general reasoning principles as possible that cannot be traced back to Mīmāṃsā sources. We extract the principles for constructing the envisioned logic directly from Mīmāṃsā texts. Since no Sanskrit philosophical school used mathematical formalization, a certain degree of abstraction is needed. The Mīmāṃsā school makes this task easier because of its insistence on using general principles of reasoning. Consequently, we can construct a logic for Maṇḍana's deontic theory solely from principles explicitly discussed or applied in relevant Mīmāṃsā texts.

The construction of such a logic requires patient, interdisciplinary teamwork. First, the rules and principles must be identified in source texts. Most of the Mīmāṃsā texts are still in Sanskrit. The source texts were translated from Sanskrit to English by our project member Elisa Freschi, a scholar of Sanskrit philosophy. After translation, these rules have to be interpreted, analyzed, and formalized. However, Mīmāṃsā authors do not always discuss reasoning principles explicitly (e.g., *nyāya*s, page 171), which means that they have to be carefully extracted from their concrete applications within Mīmāṃsā texts. Sanskrit philosophical texts usually take the form of a staged discussion among the upholders of different points of view (this approach vaguely resembles a Platonic dialogue). Consequently, inference rules are typically found within written-down discussions among several authors who invoke different rules to solve a given problem. Once identified and translated, the abstract reasoning structure underlying its concrete application must be distilled. To illustrate, from the literal translation of the *nyāya*[6] "Alternatively, [the new cloth to be used in the *mahāvrata* ritual] is additional, because it has a different purpose" together with its embedded context, we extracted a restricted version of aggregation stating that aggregation of two commands is only possible when they serve different purposes (cf. principle **P4** in Section 5.4).

Once such isolated reasoning principles are obtained, a first formalization can be provided. The resulting tentative logic enables us to derive logical consequences from these formalized principles. These logical consequences are, subsequently, compared with the use of such principles in Mīmāṃsā texts. In case of discrepancies, this often leads to a reassessment of those initial passages and the formalized logic. Furthermore, new findings in untranslated source texts may lead to modification of an earlier logical formalization. For instance, after the initial presentation of our logic for Maṇḍana's deontic theory in (van Berkel et al., 2021a), the aforementioned *nyāya* was discovered (**P4**), which led to a formal discussion of restricted aggregation in (van Berkel et al., 2022a). The logic LM is the outcome of the above interdisciplinary collaboration.

5.2 Maṇḍana's Deontic Logic of Instruments

The discussion in Section 5.1 provides the conceptual foundation of the multi-modal action logic formalizing Maṇḍana's deontic theory. We refer to this logic as the *Logic of Maṇḍana*, for short LM (Objective 1). We start by listing the central concepts we aim to capture with the formal language of LM.

Results. These are the outcomes of actions. We express results through descriptions of states of affairs, denoted by lowercase Greek letters $\varphi, \psi, \chi, \dots$. Moreover, descriptions of states of affairs may refer to actions. In those cases, the description functions as a witness to the completed performance of an action, e.g., "the Kārīri ritual has been performed".

Actions. Actions are potential instruments for obtaining results. We use $\delta, \gamma, \dots$ to represent atomic action-types, and inductively build complex action-types $\Delta, \Gamma, \dots$ (possibly indexed) using the action operators *action negation* '$-$', *disjunctive action* '$\cup$', and *joint action* '&'. (See Chapter 4 for a discussion). The performance of an *atomic* action type δ by an agent has a corresponding action token d^{δ}, which is a *propositional constant* witnessing the performance of the action by that agent. Since Maṇḍana does not deal with multi-agent interaction, the language of LM will be a single-agent language in which the agent is left implicit.

[6] *adhikaṃ vānyārthatvāt* (PMS 10.4.14).

Moments in time and the immediate future. Choice is a central property of Maṇḍana's deontic theory (cf. Section 5.4), which justifies the adaptation of an indeterministic view of time. For Maṇḍana, instrumentality statements refer to how a certain action, as an instrument, may lead to a certain state of affairs as its outcome. This 'leading to' is a temporal component referring to possible future moments. We adopt the modal operator $\boxed{\mathsf{S}}$ expressing "in all possible immediate successor moments" (some proposition holds). For example, let $\mathtt{d}^{\delta}$ stand for "the agent has thrashed the rice", then the formulae $\boxed{\mathsf{S}}\mathtt{d}^{\delta}$ is interpreted as "in all possible immediate successor moments the agent has thrashed the rice". We adopt a dual operator $\langle\mathsf{S}\rangle$ to denote "in some immediate successor moment" (some proposition holds).[7]

Universal Necessity and Facts. Although Mīmāṃsā authors (and Sanskrit philosophers in general) appeal to notions of possibility and necessity, they do not explicitly define them. We adopt a necessity modality $\boxed{\mathsf{U}}$ expressing "it is universally necessary that" (some proposition holds). We do this in order to characterize the different deontic operators better. We use statements of necessity as global assumptions, which are assertions commonly recognized as describing facts that hold in all possible situations. We refer to Blackburn et al. (2004, p.478) for a brief history of the global modality in logic.

Karma. In his deontic reduction, Maṇḍana preserves the distinction between obligatory and prohibited actions through reference to the results of these actions. In particular, fulfilling an obligation results in a reduction of bad karma (*pāpa*), whereas the result of violating a prohibition is the accumulation of bad karma. For that reason, we adopt the propositional constants $\mathtt{R}$ and $\mathtt{P}$ referring to the reduction, respectively, accumulation of bad karma. Maṇḍana's reduction is similar to the Andersonian reduction of deontic logic (1957). We discuss this in detail in Section 5.7.

Following Maṇḍana's deontic reduction, we take actions and results as the basic concepts of our language and define deontic modalities in terms of it. Our approach is similar to the one adopted in Chapter 4 and initially proposed in (van Berkel and Pascucci, 2018). First, we define two languages: an action language $\mathcal{L}^{\mathsf{Act}}$, which is an algebra of actions for agent-independent action types, and a logical language $\mathcal{L}^{\mathsf{LM}}$ into which these actions are translated. The approach allows for reasoning about complex actions from $\mathcal{L}^{\mathsf{Act}}$ as Boolean formulae in the logical language $\mathcal{L}^{\mathsf{LM}}$.

Definition 5.1 (Algebra of Actions $\mathcal{L}^{\mathsf{Act}}$)**.** *Let* $\mathsf{Act} = \{\delta, \gamma, \dots\}$ *be a non-empty countable set of* atomic action types*. The language* $\mathcal{L}^{\mathsf{Act}}$ *of complex action-types* Δ *is given via the following BNF grammar:*

$$\Delta ::= \delta \mid \Delta \cup \Delta \mid \overline{\Delta}$$

with $\delta \in \mathsf{Act}$*.*

One can see the action language $\mathcal{L}^{\mathsf{Act}}$ as a single-agent action language. We define *joint action* & in terms of action negation and disjunctive action, i.e., $\Delta \& \Gamma := \overline{\overline{\Delta} \cup \overline{\Gamma}}$.

[7] In contrast to the Logic of Action and Norms in Chapter 4, we do not require explicit reference to the actual future in defining Maṇḍana's reduction.

In what follows, we employ a reductionist approach to deontic concepts via "karma constants" and a reduction of actions via action constants (cf. Chapter 4). Let $\mathsf{Wit} = \{\mathtt{d}^{\delta}, \mathtt{d}^{\gamma}, \dots\}$ be the set of propositional constants that witness the performance of atomic action-types $\delta, \gamma, \dots \in \mathsf{Act}$ by the agent in question. We take $\mathtt{d}^{\delta}$ to read "the agent has performed action δ". The formal correspondence between agent-dependent action types and propositional constants is given in Definition 5.3 below. Furthermore, let $\mathtt{P}$ and $\mathtt{R}$ be propositional constants witnessing "bad karma is accumulated" and "bad karma is reduced", respectively.

Definition 5.2 (The Language $\mathcal{L}^{\mathsf{LM}}$)**.** *Let* $\mathsf{Atoms} = \{p, q, r, \dots\}$ *be a countable set of atomic propositions. The language* $\mathcal{L}^{\mathsf{LM}}$ *is given by the following BNF:*

$$\varphi ::= p \mid \mathtt{d}^{\delta} \mid \mathtt{P} \mid \mathtt{R} \mid \neg\varphi \mid \varphi \vee \varphi \mid \boxed{\mathsf{S}}\varphi \mid \boxed{\mathsf{U}}\varphi$$

where $p \in \mathsf{Atoms}$ *and* $\mathtt{d}^{\delta} \in \mathsf{Wit}$*.*

The other connectives $\wedge$ and $\rightarrow$, as well as $\top$ and $\bot$, are defined as usual. Formulae of the form $\boxed{\mathsf{S}}\varphi$ and $\boxed{\mathsf{U}}\varphi$ express, respectively, "in all possible immediate successor moments φ holds" and "it is universally necessary that φ holds". We sometimes omit reference to 'immediate' in referring to $\boxed{\mathsf{S}}$. We take $\langle\mathsf{S}\rangle$ and $\langle\mathsf{U}\rangle$ as the duals of $\boxed{\mathsf{S}}$ and $\boxed{\mathsf{U}}$, respectively.

Definition 5.3 (Translation between $\mathcal{L}^{\mathsf{Act}}$ and $\mathcal{L}^{\mathsf{LM}}$)**.** *The translation t encoding action-types from* $\mathcal{L}^{\mathsf{Act}}$ *into formulae of* $\mathcal{L}^{\mathsf{LM}}$ *is established recursively:*

- *For any* $\delta \in \mathsf{Act}$, $t(\delta) = \mathtt{d}^{\delta}$, *with* $\mathtt{d}^{\delta} \in \mathcal{L}^{\mathsf{LM}}$;
- *For any* $\Delta \in \mathcal{L}^{\mathsf{Act}}$, $t(\overline{\Delta}) = \neg t(\Delta)$;
- *For any* $\Delta, \Gamma \in \mathcal{L}^{\mathsf{Act}}$, $t(\Delta \cup \Gamma) = t(\Delta) \vee t(\Gamma)$.

As an example of a formula from the language $\mathcal{L}^{\mathsf{LM}}$, consider

$$[\Delta]\varphi := \boxed{\mathsf{S}}(t(\Delta) \rightarrow \varphi)$$

which reads "at every successor world witnessing the performance of action Δ, the state of affairs φ holds". The language $\mathcal{L}^{\mathsf{LM}}$ is similar to the one employed in Chapter 4, and we refer to that chapter for an extensive discussion of its expressivity (e.g., in relation to PDL).

5.2.1 Axiomatization of LM

We use classical propositional logic as our base logic. The use of classical logic instead of, for instance, intuitionistic logic as adopted in (Abraham et al., 2011), is motivated by various examples found in Mīmāṃsā texts which implicitly assume the legitimacy of excluded middle and reductio ad absurdum (Ciabattoni et al., 2015). To illustrate, consider the following Mīmāṃsā principle from Jayanta's book *Nyāyamañjarī*: "When there is a contradiction, at the denial of one [alternative], the other is known [to be true]". The Hilbert-style axiomatization of the logic LM is given below.

Definition 5.4 (The Axiomatization of LM)**.** *The logic* LM *is axiomatized by the following collection of axiom schemes and rules:*

- $\mathsf{A0}$. *All classical propositional tautologies;*
- $\mathsf{R0}$. *From* φ *and* $\varphi \rightarrow \psi$*, infer* ψ*;*
- $\mathsf{A1}$. $\boxed{\mathsf{U}}(\varphi \rightarrow \psi) \rightarrow (\boxed{\mathsf{U}}\varphi \rightarrow \boxed{\mathsf{U}}\psi)$*;*
- $\mathsf{A2}$. $\boxed{\mathsf{U}}\varphi \rightarrow \varphi$*;*
- $\mathsf{A3}$. $\Diamond_{\mathsf{U}}\varphi \rightarrow \boxed{\mathsf{U}}\Diamond_{\mathsf{U}}\varphi$*;*
- $\mathsf{A4}$. $\boxed{\mathsf{S}}(\varphi \rightarrow \psi) \rightarrow (\boxed{\mathsf{S}}\varphi \rightarrow \boxed{\mathsf{S}}\psi)$*;*
- $\mathsf{A5}$. $\boxed{\mathsf{U}}\varphi \rightarrow \boxed{\mathsf{S}}\varphi$*;*
- $\mathsf{A6}$. $\Diamond_{\mathsf{S}}\mathtt{P} \rightarrow \Diamond_{\mathsf{S}}\neg\mathtt{P}$*;*
- $\mathsf{A7}$. $\Diamond_{\mathsf{S}}\mathtt{R} \rightarrow \Diamond_{\mathsf{S}}\neg\mathtt{R}$*;*
- $\mathsf{R1}$. *From* φ*, infer* $\boxed{\mathsf{U}}\varphi$*;*

The logic LM *is the smallest set of formulae from* $\mathcal{L}^{\mathsf{LM}}$ *closed under all instances of the axiom schemes, and applications of the inference rules* $\mathsf{R0}$ – $\mathsf{R1}$*. Whenever* $\varphi \in \mathsf{LM}$ *we say that* $\varphi \in \mathcal{L}^{\mathsf{LM}}$ *is a* LM-theorem *and write* $\vdash_{\mathsf{LM}} \varphi$*. Last,* LM*-derivability is defined as usual (Definition 2.3).*

The axiomatization of LM is deliberately minimal, i.e., all properties except for those related to the universal necessity modality $\boxed{\mathsf{U}}$ can be traced back to Maṇḍana (see page 175 for the motivation). Furthermore, we emphasize that it suffices to adopt a general notion of the immediate successor modality $\boxed{\mathsf{S}}$ since Maṇḍana's analysis does not depend on inherent properties of time.[8] Both $\boxed{\mathsf{U}}$ and $\boxed{\mathsf{S}}$ are normal modal operators by virtue of $\mathsf{A1}$, $\mathsf{A4}$, $\mathsf{A5}$, and $\mathsf{R1}$. Axioms $\mathsf{A2}$ and $\mathsf{A3}$ characterize $\boxed{\mathsf{U}}$ as an $\mathsf{S5}$ modality. Furthermore, $\mathsf{A5}$ is a bridge axiom, expressing that what holds universally must also hold at any successor moment.

The two Maṇḍana inspired axioms are $\mathsf{A6}$ and $\mathsf{A7}$. The former expresses that if there is an immediate successor in which bad karma is accumulated, there is also a successor moment in which it can be avoided. The latter expresses similar reasoning but then concerns the reduction of bad karma. Both axioms are based on a Mīmāṃsā principle, endorsed by Maṇḍana, which states that all commands must be *non-trivial* (Freschi, 2018). To see this point, suppose towards a contradiction that at all successor moments bad karma is accumulated, then whatever the agent does, bad karma will be obtained.

[8] One may refine the immediate successor modality by additionally imposing, e.g., intransitivity and asymmetry (cf. Chapter 2 on temporal STIT logic). This goes beyond our objective in this chapter.

Consequently, each corresponding command at the moment will be trivially violated. This conflicts with the above Mīmāṃsā principle. The same reasoning applies to the reduction of bad karma.

We point out that, from the perspective of the axiomatization, the constants $\mathtt{P}$ and $\mathtt{R}$ show the same logical behavior. Only in defining obligations and prohibitions do these constants receive a different meaning, with prohibited actions leading to an accumulation and obligatory actions to a reduction of bad karma. We come back to this in Section 5.4. The use of constants in characterizing certain properties of the logic of Maṇḍana is similar to the approach by Anderson and Moore (1957). We compare these approaches in Section 5.7.

Remark 5.1. *The logic* LM *was first presented in (van Berkel et al., 2022a) and differs from the Maṇḍana logic* LMa *in (van Berkel et al., 2021a). The latter includes the additional axiom* $\langle\mathsf{S}\rangle t(\Delta) \rightarrow \langle\mathsf{S}\rangle(t(\Delta) \wedge (\neg\mathtt{R} \vee \neg\mathtt{P}))$. *The modification was motivated by the formalization of a Mīmāṃsā principle discovered after the publication of the latter (cf.* ***P4*** *on page 190). We refer to (van Berkel et al., 2022a) for a comparison of* LM *and* LMa*.*

The logic LM does not fully represent Maṇḍana's deontic theory but contains its essential building blocks.[9] That is, we will formally *define* the central concepts of Maṇḍana's theory, using the immediate successor modality, complex action types, and karma constants. We do this in Section 5.4.

5.2.2 Semantics for LM

We use relational semantics to characterize LM (Blackburn et al., 2004).

Definition 5.5 (Frames and Models for LM)**.** *An* LM-frame *is defined as a tuple* $\mathfrak{F} = \langle W, \{W_{\mathtt{d}^\delta} \mid \mathtt{d}^\delta \in \mathcal{L}^{\mathsf{LM}}\}, W_{\mathtt{P}}, W_{\mathtt{R}}, \mathcal{R}_{\boxed{\mathsf{U}}}, \mathcal{R}_{\boxed{\mathsf{S}}}\rangle$. *Let* W *be a non-empty set of worlds* $w, v, u, \ldots$, *let* $\mathcal{R}_{\boxed{\mathsf{U}}} = W \times W$, $\mathcal{R}_{\boxed{\mathsf{S}}} \subseteq W \times W$, *and* $\mathcal{R}_{[\alpha]}(w) := \{v \in W \mid (w, v) \in \mathcal{R}_{[\alpha]}\}$ *for* $[\alpha] \in \{\boxed{\mathsf{U}}, \boxed{\mathsf{S}}\}$. *The following hold:*

- ***R1*** *For each* $\mathtt{d}^\delta \in \mathsf{Wit}$, $W_{\mathtt{d}^\delta} \subseteq W$*;*
- ***R2*** $W_{\mathtt{P}} \subseteq W$*;*
- ***R3*** $W_{\mathtt{R}} \subseteq W$*;*
- ***R4*** *For all* $w, v \in W$, *if* $v \in \mathcal{R}_{\boxed{\mathsf{S}}}(w)$ *and* $v \in W_{\mathtt{P}}$, *then there is a* $u \in W$ *such that* $u \in \mathcal{R}_{\boxed{\mathsf{S}}}(w)$ *and* $u \notin W_{\mathtt{P}}$*;*

[9] We stress that the logic LM is developed for reasoning *about* Vedic commands as interpreted by Maṇḍana. All Mīmāṃsā authors consider Vedic commands to be self-contained and immutable. This means that no new Vedic command can be derived through logic. Thus, the logic LM is used to derive deontic consequences from Vedic commands and deals with commands on the *derived level* instead of the Vedic level.

R5 *For all $w, v \in W$, if $v \in \mathcal{R}_{\boxed{\mathsf{S}}}(w)$ and $v \in W_{\mathtt{R}}$, then there is a $u \in W$ such that $u \in \mathcal{R}_{\boxed{\mathsf{S}}}(w)$ and $u \notin W_{\mathtt{R}}$.*

An LM*-model is a tuple $\mathfrak{M} = \langle \mathfrak{F}, V \rangle$ where $\mathfrak{F}$ is an* LM*-frame and V is a valuation function mapping propositional atoms and constants to subsets of W, i.e., $V : \mathsf{Atoms} \cup \mathsf{Wit} \cup \{\mathtt{P}, \mathtt{R}\} \mapsto \mathcal{P}(W)$, for which the following three restrictions hold:*

- *$V(\mathtt{d}^{\delta}) = W_{\mathtt{d}^{\delta}}$ for any $\mathtt{d}^{\delta} \in \mathsf{Wit}$;*
- *$V(\mathtt{P}) = W_{\mathtt{P}}$;*
- *$V(\mathtt{R}) = W_{\mathtt{R}}$.*

In Definition 5.5, condition **R1** stipulates that the set $W_{\mathtt{d}^{\delta}}$ (for each $\mathtt{d}^{\delta} \in \mathsf{Wit}$) contains those moments from W witnessing the successful performance of δ. The first restriction on the valuation function V ensures that those moments witnessing $\mathtt{d}^{\delta}$ satisfy those constants. That this is also the case for arbitrary actions is shown in Lemma 5.1. Conditions **R2** and **R3** ensure that the sets $W_{\mathtt{P}} \subseteq W$ and $W_{\mathtt{R}} \subseteq W$ contain those moments from W witnessing the accumulation, respectively reduction of bad karma. Hence, in LM-models, the valuation of constants is fixed on the level of LM-frames. This means that the semantic interpretation of such constants is fixed for every model defined over that frame. The frame properties **R4** and **R5** make use of this fact. **R4** conveys that whenever bad karma is attainable, it is also avoidable (cf. A6), whereas **R5** captures the same property for the reduction of bad karma (cf. A7). Last, we point out that the $\boxed{\mathsf{U}}$-modality represents the global modality and is therefore characterized through an equivalence relation whose equivalence class is the set of all worlds W, i.e., $\mathcal{R}_{\boxed{\mathsf{U}}} = W \times W$. Consequently, since $\mathcal{R}_{\boxed{\mathsf{S}}} \subseteq W \times W$, any $\boxed{\mathsf{S}}$ transition is also a $\boxed{\mathsf{U}}$ transition (cf. A5).

The semantic interpretation of $\mathcal{L}^{\mathsf{LM}}$ is defined below.

Definition 5.6 (Semantics of LM-models)**.** *Let $\mathfrak{M}$ be an* LM*-model and let $w \in W$ of $\mathfrak{M}$. The* satisfaction *of a formula $\varphi \in \mathcal{L}^{\mathsf{LM}}$ in $\mathfrak{M}$ at w is defined accordingly:*

1. *$\mathfrak{M}, w \models p$ iff $w \in V(p)$;*
2. *$\mathfrak{M}, w \models \chi$ iff $w \in V(\chi) = W_{\chi}$ for any $\chi \in \mathsf{Wit} \cup \{\mathtt{P}, \mathtt{R}\}$;*
3. *$\mathfrak{M}, w \models \neg\varphi$ iff not $\mathfrak{M}, w \models \varphi$;*
4. *$\mathfrak{M}, w \models \varphi \vee \psi$ iff $\mathfrak{M}, w \models \varphi$ or $\mathfrak{M}, w \models \psi$;*
5. *$\mathfrak{M}, w \models \boxed{\mathsf{U}}\varphi$ iff for all $v \in \mathcal{R}_{\boxed{\mathsf{U}}}(w)$, $\mathfrak{M}, v \models \varphi$;*
6. *$\mathfrak{M}, w \models \boxed{\mathsf{S}}\varphi$ iff for all $v \in \mathcal{R}_{\boxed{\mathsf{S}}}(w)$, $\mathfrak{M}, v \models \varphi$.*

Global truth, validity, and semantic entailment are defined as usual (Definition 2.5).

In Definition 5.5, we restricted the valuation of action constants to sets of worlds. The lemma below shows that this observation can be generalized to arbitrary action types.

Lemma 5.1. *For an arbitrary* $\Delta \in \mathcal{L}^{\mathsf{Act}}$*, we define* $W_{t(\Delta)}$ *using the following recursive clauses:* $W_{t(\delta)} := W_{\mathsf{d}^{\delta}}$*;* $W_{t(\overline{\Delta})} := W \setminus W_{t(\Delta)}$*;* $W_{t(\Delta \cup \Gamma)} := W_{t(\Delta)} \cup W_{t(\Gamma)}$*. For each* LM*-model* $\mathfrak{M}$ *and each moment* $w \in W$ *of* $\mathfrak{M}$*, we have:* $\mathfrak{M}, w \models t(\Delta)$ *iff* $w \in W_{t(\Delta)}$*.*

Proof. The proof is the same as for Lemma 4.1 in Chapter 4. QED

5.3 Soundness and Completeness

Soundness is demonstrated as usual, and completeness is shown through a modification of the canonical model approach. The modification is required due to the universal necessity modality in LM.

Theorem 5.1 (Soundness of LM)**.** *For any formula* $\varphi \in \mathcal{L}^{\mathsf{LM}}$*, and any* $\Gamma \subseteq \mathcal{L}^{\mathsf{LM}}$*: if* $\Gamma \vdash_{\mathsf{LM}} \varphi$*, then* $\Gamma \models_{\mathsf{LM}} \varphi$*.*

Proof. It suffices to show that all axioms are LM-valid. Take an arbitrary LM-model $\mathfrak{M}$ and an arbitrary $w \in W$ of $\mathfrak{M}$. The axiom schemes A0, A1, and A4, and rules R0 and R1 are valid, respectively preserve validity on all relational frames (Blackburn et al., 2004). We omit their proofs.

A2 $\boxed{\mathsf{U}}\varphi \rightarrow \varphi$. Assume $\mathfrak{M}, w \models \boxed{\mathsf{U}}\varphi$. By the semantic definition of $\boxed{\mathsf{U}}$, for all $v \in W$, $\mathfrak{M}, v \models \varphi$ and since $w \in W$ we have $\mathfrak{M}, w \models \varphi$.

A3 $\langle\mathsf{U}\rangle\varphi \rightarrow \boxed{\mathsf{U}}\langle\mathsf{U}\rangle\varphi$. Assume $\mathfrak{M}, w \models \langle\mathsf{U}\rangle\varphi$. By semantic definition of $\langle\mathsf{U}\rangle$, there is a $v \in W$ such that $\mathfrak{M}, v \models \varphi$. Now, take an arbitrary $u \in W$. Since both $u, v \in W$ we have that $\mathfrak{M}, u \models \langle\mathsf{U}\rangle\varphi$ and since u is arbitrary we have $\mathfrak{M}, w \models \boxed{\mathsf{U}}\langle\mathsf{U}\rangle\varphi$.

A5 $\boxed{\mathsf{U}}\varphi \rightarrow \boxed{\mathsf{S}}\varphi$. Assume $\mathfrak{M}, w \models \boxed{\mathsf{U}}\varphi$. By semantic definition of $\boxed{\mathsf{U}}$ we know that for all $v \in W$, $\mathfrak{M}, v \models \varphi$. By the fact that $\mathcal{R}_{\boxed{\mathsf{S}}}(w) \subseteq W$ we know that for all $v \in \mathcal{R}_{\boxed{\mathsf{S}}}(w)$, $\mathfrak{M}, v \models \varphi$ too. Hence, $\mathfrak{M}, w \models \boxed{\mathsf{S}}\varphi$.

A6 $\langle\mathsf{S}\rangle\mathtt{P} \rightarrow \langle\mathsf{S}\rangle\neg\mathtt{P}$. Assume $\mathfrak{M}, w \models \langle\mathsf{S}\rangle\mathtt{P}$. By semantic definition of $\langle\mathsf{S}\rangle$ there is a $v \in W$ such that $v \in \mathcal{R}_{\boxed{\mathsf{S}}}(w)$ with $\mathfrak{M}, v \models \mathtt{P}$. Hence, $v \in W_{\mathtt{P}}$. By **R4** Definition 5.5, we know that there is a $u \in W$ such that $u \in \mathcal{R}_{\boxed{\mathsf{S}}}(w)$ and $u \notin W_{\mathtt{P}}$. Therefore, $\mathfrak{M}, u \models \neg\mathtt{P}$ and so $\mathfrak{M}, w \models \langle\mathsf{S}\rangle\neg\mathtt{P}$.

A7 Similar to A6.

The main claim follows through reasoning similar to Theorem 4.1 of Section 4.3. QED

Completeness is proven via a canonical model construction. Due to the universal necessity modality $\boxed{\mathsf{U}}$ we must modify the standard completeness by canonicity strategy (Blackburn et al., 2004). Recall that $\boxed{\mathsf{U}}$ is axiomatized as an $\mathsf{S5}$ modality, which is canonical for the equivalence relation $\mathcal{R}^c_{\boxed{\mathsf{U}}}$ (Blackburn et al., 2004, Ch.4). Using the standard canonical model construction (as in Chapter 4) only guarantees that $\mathcal{R}^c_{\boxed{\mathsf{U}}} \subseteq W \times W$ but not that $\mathcal{R}^c_{\boxed{\mathsf{U}}} = W \times W$, as desired. Namely, the $\mathsf{S5}$ characterization of $\boxed{\mathsf{U}}$ may generate several $\mathcal{R}^c_{\boxed{\mathsf{U}}}$-equivalences classes in the canonical model. The solution is as follows (Blackburn et al., 2004, Ch.7): we use generated submodels of the canonical model and prove that these are LM-models. The submodels are defined relative to a given LM-maximally consistent set Σ and denoted by $\mathfrak{M}^{\Sigma}$.

We first provide the usual preliminaries. The proofs are omitted (e.g., see Section 4.3). Notice that we reserved $\Delta, \Gamma, \dots$ for arbitrary action types of $\mathcal{L}^{\mathsf{Act}}$. In order to enhance clarity in the completeness proof, we use $\Sigma, \Theta, \Pi, \dots$ to refer to LM-MCSs.

Definition 5.7 (LM-CS and LM-MCS)**.** *A set $\Sigma \subset \mathcal{L}^{\mathsf{LM}}$ is a* LM-consistent *(*LM*-CS) iff $\Sigma \nvdash_{\mathsf{LM}} \bot$. A set $\Sigma \subset \mathcal{L}^{\mathsf{LM}}$ is an* LM-maximally consistent *(*LM*-MCS) iff Σ is an* LM*-CS and for any set $\Sigma' \subseteq \mathcal{L}^{\mathsf{LM}}$ such that $\Sigma \subset \Sigma'$ it is the case that $\Sigma' \vdash_{\mathsf{LM}} \bot$.*

In what follows, we make use of the standard properties of MCSs. See Section 2.2 for proofs. We use these properties implicitly throughout the section.

Lemma 5.2 (Properties of MCSs)**.** *Let $\Sigma \subseteq \mathcal{L}^{\mathsf{LM}}$ be an* LM*-MCS and $\varphi \in \mathcal{L}^{\mathsf{LM}}$. The following holds:*

- *$\Sigma \vdash_{\mathsf{LM}} \varphi$ iff $\varphi \in \Sigma$;*
- *$\varphi \in \Sigma$ iff $\neg\varphi \notin \Sigma$;*
- *$\varphi \wedge \psi \in \Sigma$ iff $\varphi \in \Sigma$ and $\psi \in \Sigma$.*

Adapting Lindenbaum's Lemma to the context of LM, we know that every LM-CS can be extended to an LM-MCS.

Lemma 5.3 (Lindenbaum's Lemma)**.** *Let $\Sigma \subseteq \mathcal{L}^{\mathsf{LM}}$ be an* LM*-CS: there is an* LM*-MCS $\Sigma' \subseteq \mathcal{L}^{\mathsf{LM}}$ such that $\Sigma \subseteq \Sigma'$.*

Definition 5.8 (Canonical model for LM)**.** *We define the* canonical model $\mathfrak{M}^c$ *to be the tuple $\mathfrak{M}^c = \langle W^c, \{W^c_{\mathsf{d}^\delta} \mid \mathsf{d}^\delta \in \mathsf{Wit}\}, W^c_{\mathtt{P}}, W^c_{\mathtt{R}}, \mathcal{R}^c_{\boxed{\mathsf{U}}}, \mathcal{R}^c_{\boxed{\mathsf{S}}}, V^c\rangle$ such that:*

- *$W^c := \{\Sigma \subset \mathcal{L}^{\mathsf{LM}} \mid \Sigma$ is a* LM*-MCS$\}$;*
- *For all $\mathsf{d}^\delta \in \mathsf{Wit}$, $W^c_{\mathsf{d}^\delta} := \{\Sigma \in W^c \mid \mathsf{d}^\delta \in \Sigma\}$;*
- *For each $\alpha \in \{\mathtt{P}, \mathtt{R}\}$, $W^c_\alpha := \{\Sigma \in W^c \mid \alpha \in \Sigma\}$;*
- *For each $\Sigma \in W^c$, $\mathcal{R}^c_{\boxed{\mathsf{U}}}(\Sigma) := \{\Theta \in W^c \mid$ for all $\boxed{\mathsf{U}}\varphi \in \Sigma$, $\varphi \in \Theta\}$;*

- *For each* $\Sigma \in W^c$, $\mathcal{R}^c_{\boxed{\mathtt{S}}}(\Sigma) := \{\Theta \in W^c \mid$ *for all* $\boxed{\mathtt{S}}\varphi \in \Sigma$, $\varphi \in \Theta\}$;
- V^c *is a valuation function such that for all* $\chi \in \mathsf{Atoms} \cup \mathsf{Wit} \cup \{\mathtt{P}, \mathtt{R}\}$, $V^c(\chi) := \{\Sigma \in W^c \mid \chi \in \Sigma\}$.

The semantic evaluation of $\mathcal{L}^{\mathsf{LM}}$ *formulae on* $\mathfrak{M}^c$ *is defined as in Definition 5.6.*

The canonical model possesses the usual properties (see Section 2.2 for the proofs).

Lemma 5.4 (Existence Lemma $\boxed{\mathtt{U}}$ and $\boxed{\mathtt{S}}$)**.** *For any world* $\Sigma \in W^c$ *of* $\mathfrak{M}^c$ *the following holds:*

- *If* $\langle\mathtt{U}\rangle\varphi \in \Sigma$, *then there is a* $\Theta \in W^c$ *such that* $\varphi \in \Theta$ *and* $\Theta \in \mathcal{R}_{\boxed{\mathtt{U}}}(\Sigma)$;
- *If* $\langle\mathtt{S}\rangle\varphi \in \Sigma$, *then there is a* $\Theta \in W^c$ *such that* $\varphi \in \Theta$ *and* $\Theta \in \mathcal{R}_{\boxed{\mathtt{S}}}(\Sigma)$.

Corollary 5.1. *For any world* $\Sigma \in W^c$ *of* $\mathfrak{M}^c$ *the following holds:*

- *If for all* $\Theta \in \mathcal{R}_{\boxed{\mathtt{U}}}(\Sigma), \varphi \in \Theta$, *then* $\boxed{\mathtt{U}}\varphi \in \Sigma$;
- *If for all* $\Theta \in \mathcal{R}_{\boxed{\mathtt{S}}}(\Sigma), \varphi \in \Theta$, *then* $\boxed{\mathtt{S}}\varphi \in \Sigma$.

The following lemma shows that the defined model is canonical for the logic LM, i.e., each LM-MCS is satisfiable on this model.

Lemma 5.5 (Truth Lemma)**.** *For any* $\varphi \in \mathcal{L}^{\mathsf{LM}}$ *and* $\Sigma \in W^c$*:* $\mathfrak{M}^c, \Sigma \models \varphi$ *iff* $\varphi \in \Sigma$.

As discussed, the model $\mathfrak{M}^c$ is not necessarily an LM-model. In Definition 5.9 we define submodels of $\mathfrak{M}^c$ that do belong to the class of LM-models (Lemma 5.8).

Definition 5.9 (Submodel of the Canonical Model for LM)**.** *Let* $\mathfrak{M}^c$ *be the canonical model for* LM*. We define the* submodel $\mathfrak{M}^\Sigma$ *relative to* $\Sigma \in W^c$ *to be the tuple* $\mathfrak{M}^\Sigma = \langle W^\Sigma, \{W^\Sigma_{\mathsf{d}^\delta} \mid \mathsf{d}^\delta \in \mathsf{Wit}\}, W^\Sigma_{\mathtt{P}}, W^\Sigma_{\mathtt{R}}, \mathcal{R}^\Sigma_{\boxed{\mathtt{U}}}, \mathcal{R}^\Sigma_{\boxed{\mathtt{S}}}, V^\Sigma\rangle$ *such that:*

- $W^\Sigma := \{\Theta \subset \mathcal{L}^{\mathsf{LM}} \mid \Theta$ *is a* LM*-MCS and for each* $\boxed{\mathtt{U}}\varphi \in \Sigma, \varphi \in \Theta\}$;
- *For all* $\mathsf{d}^\delta \in \mathsf{Wit}$, $W^\Sigma_{\mathsf{d}^\delta} := W^c_{\mathsf{d}^\delta} \cap W^\Sigma$;
- *For each* $\alpha \in \{\mathtt{P}, \mathtt{R}\}$, $W^\Sigma_\alpha := W^c_\alpha \cap W^\Sigma$;
- *For each* $[\alpha] \in \{\boxed{\mathtt{U}}, \boxed{\mathtt{S}}\}$ *and all* $\Theta \in W^\Sigma$, $\mathcal{R}^\Sigma_{[\alpha]}(\Theta) := \mathcal{R}^c_{[\alpha]}(\Theta) \cap W^\Sigma$;
- V^Σ *is a valuation function such that for all* $\chi \in \mathsf{Atoms} \cup \mathsf{Wit} \cup \{\mathtt{P}, \mathtt{R}\}$, $V^\Sigma(\chi) := V^c(\chi) \cap W^\Sigma$.

First, by the definition of W^{Σ} and $\mathcal{R}^{\Sigma}_{\boxed{\mathsf{U}}}$, it can be immediately seen that $\mathfrak{M}^{\Sigma}$ satisfies the condition $\mathcal{R}^{\Sigma}_{\boxed{\mathsf{U}}} = W^{\Sigma} \times W^{\Sigma}$. Furthermore, the model $\mathfrak{M}^{\Sigma}$ is a *generated submodel* of $\mathfrak{M}^{c}$, this follows from Lemma 5.6. We refer to Blackburn et al. (2004, Ch.2) for general results on generated submodels.

Lemma 5.6. *Let* $[\alpha] \in \{\boxed{\mathsf{U}}, \boxed{\mathsf{S}}\}$*. For each* $\Theta, \Pi \in W^{c}$ *of* $\mathfrak{M}^{c}$*, if* $\Theta \in W^{\Sigma}$ *and* $\Pi \in \mathcal{R}^{c}_{[\alpha]}(\Theta)$*, then* $\Pi \in W^{\Sigma}$*.*

Proof. Assume $\Theta \in W^{\Sigma}$ and $\Pi \in \mathcal{R}^{c}_{[\alpha]}(\Theta)$. Suppose towards a contradiction that $\Pi \notin W^{\Sigma}$. By construction of $\mathfrak{M}^{\Sigma}$, there is a $\boxed{\mathsf{U}}\varphi \in \Sigma$ such that $\varphi \notin \Pi$. Since $\Theta \in W^{\Sigma}$ we know $\boxed{\mathsf{U}}\varphi \in \Theta$. Since Θ is an LM-MCS we have $\boxed{\mathsf{U}}\varphi \to \boxed{\mathsf{S}}\varphi \in \Theta$ and so both $\boxed{\mathsf{S}}\varphi, \boxed{\mathsf{U}}\varphi \in \Theta$. By construction of the canonical model $\mathfrak{M}^{c}$ and the assumption that $\Pi \in \mathcal{R}^{c}_{[\alpha]}(\Theta)$ we have $\varphi \in \Pi$. Contradiction. QED

Lemma 5.7. *For each* $\varphi \in \mathcal{L}^{\mathsf{LM}}$ *and each* $\Theta \in W^{\Sigma}$ *of* $\mathfrak{M}^{\Sigma}$*:* $\mathfrak{M}^{c}, \Theta \models \varphi$ *iff* $\mathfrak{M}^{\Sigma}, \Theta \models \varphi$*.*

Proof. The proof is by induction on the complexity of φ. *Base case.* $\varphi = \chi \in \mathsf{Atoms} \cup \mathsf{Wit} \cup \{\mathtt{P}, \mathtt{R}\}$. Trivial since $\Theta \in V^{\Sigma}(\chi) = V^{c} \cap W^{\Sigma}$. *Inductive step.* We only consider the modal case $\varphi = \boxed{\mathsf{U}}\psi$, the proof of $\varphi = \boxed{\mathsf{S}}\psi$ is similar. Left-to-Right. Assume $\mathfrak{M}^{c}, \Theta \models \boxed{\mathsf{U}}\psi$. By semantic definition, for all $\Pi \in \mathcal{R}^{c}_{\boxed{\mathsf{U}}}(\Theta)$ we have $\mathfrak{M}^{c}, \Pi \models \psi$. By Lemma 5.6 we know that for all $\Pi \in \mathcal{R}^{c}_{\boxed{\mathsf{U}}}(\Theta), \Pi \in W^{\Sigma}$, and by the IH we obtain $\mathfrak{M}^{\Sigma}, \Pi \models \psi$. Since $\mathcal{R}^{\Sigma}_{\boxed{\mathsf{U}}}(\Theta) = \mathcal{R}^{c}_{\boxed{\mathsf{U}}}(\Theta) \cap W^{\Sigma}$, we have $\mathfrak{M}^{\Sigma}, \Theta \models \boxed{\mathsf{U}}\psi$. Right-to-Left. We prove this by contraposition. Assume $\mathfrak{M}^{c}, \Theta \not\models \boxed{\mathsf{U}}\psi$. By semantic definition there is a $\Pi \in \mathcal{R}^{c}_{\boxed{\mathsf{U}}}(\Theta)$ such that $\mathfrak{M}^{c}, \Pi \not\models \psi$. By Lemma 5.6 we know $\Pi \in W^{\Sigma}$ too. Hence, by IH $\mathfrak{M}^{\Sigma}, \Pi \not\models \psi$. Since $\Pi, \Theta \in W^{\Sigma}$ we have $\Pi \in \mathcal{R}^{c}_{\boxed{\mathsf{U}}}(\Theta) \cap W^{\Sigma} = \mathcal{R}^{\Sigma}_{\boxed{\mathsf{U}}}$ and so $\mathfrak{M}^{\Sigma}, \Theta \not\models \boxed{\mathsf{U}}\psi$. QED

We show that $\mathfrak{M}^{\Sigma}$ belongs to the class of LM-models.

Lemma 5.8 (LM-submodel)**.** *The submodel* $\mathfrak{M}^{\Sigma}$ *generated from the canonical model* $\mathfrak{M}^{c}$ *is an* LM*-model.*

Proof. It can be easily observed that W^{Σ} and V^{Σ} (for $\chi \in \mathsf{Atoms} \cup \mathsf{Wit} \cup \{\mathtt{P}, \mathtt{R}\}$) are well-defined. Note, that W^{Σ} is non-empty since by $\boxed{\mathsf{U}}\varphi \to \varphi \in \Sigma$ and Definition 5.9, it is the case that $\Sigma \in W^{\Sigma}$ (that W^{c} is non-empty follows from the model defined in Figure 5.2 Section 5.5). We only need to show that $\mathfrak{M}^{\Sigma}$ satisfies the properties **R1**–**R5**.

R1 For each $\mathsf{d}^{\delta} \in \mathsf{Wit}$, $W^{c}_{\mathsf{d}^{\delta}} \subseteq W^{c}$ follows directly from the definition of $W^{c}_{\mathsf{d}^{\delta}}$.

R2 Similar to **R1**.

R3 Similar to **R1**.

R4 Take arbitrary $\Theta, \Pi \in W^{\Sigma}$ and assume $\Pi \in \mathcal{R}^{\Sigma}_{\boxed{\mathsf{S}}}(\Theta)$ and $\Pi \in W^{\Sigma}_{\mathsf{P}}$. We construct a LM-MCS Ω such that $\Omega \in \mathcal{R}^{\Sigma}_{\boxed{\mathsf{S}}}(\Theta)$ and $\Omega \notin W^{\Sigma}_{\mathsf{P}}$. Let,

$$\Omega' = \{\neg \mathsf{P}\} \cup \{\varphi \mid \boxed{\mathsf{S}}\varphi \in \Theta\} \cup \{\psi \mid \boxed{\mathsf{U}}\psi \in \Theta\}$$

Suppose towards a contradiction that Ω' is not LM-consistent. Hence for some $\varphi_1, .., \varphi_n, \psi_1, \ldots, \psi_m \in \Omega'$, we have $\vdash_{\mathsf{LM}} (\varphi \wedge \cdots \wedge \varphi_n \wedge \psi_1 \wedge \cdots \wedge \psi_m) \to \mathsf{P}$. By the normality of $\boxed{\mathsf{S}}$, we have $\vdash_{\mathsf{LM}} \boxed{\mathsf{S}}((\varphi \wedge \cdots \wedge \varphi_n \wedge \psi_1 \wedge \cdots \wedge \psi_m) \to \mathsf{P})$, which implies $\vdash_{\mathsf{LM}} \boxed{\mathsf{S}}(\varphi \wedge \cdots \wedge \varphi_n \wedge \psi_1 \wedge \cdots \wedge \psi_m) \to \boxed{\mathsf{S}}\mathsf{P}$. By normality of $\boxed{\mathsf{S}}$ and monotonicity of LM, $\vdash_{\mathsf{LM}} (\boxed{\mathsf{S}}\varphi \wedge \cdots \wedge \boxed{\mathsf{S}}\varphi_n \wedge \boxed{\mathsf{S}}\psi_1 \wedge \cdots \wedge \boxed{\mathsf{S}}\psi_m \wedge \langle\mathsf{S}\rangle\mathsf{P}) \to \neg\langle\mathsf{S}\rangle\neg\mathsf{P}$ and by maximal consistency of Θ, $\vdash_{\mathsf{LM}} (\boxed{\mathsf{S}}\varphi \wedge \cdots \wedge \boxed{\mathsf{S}}\varphi_n \wedge \boxed{\mathsf{S}}\psi_1 \wedge \cdots \wedge \boxed{\mathsf{S}}\psi_m \wedge \langle\mathsf{S}\rangle\mathsf{P}) \to \neg\langle\mathsf{S}\rangle\neg\mathsf{P} \in \Theta$. By assumption $\boxed{\mathsf{S}}\varphi_1, \ldots, \boxed{\mathsf{S}}\varphi_n, \langle\mathsf{S}\rangle\mathsf{P} \in \Theta$. Furthermore, by assumption $\boxed{\mathsf{U}}\psi_i \in \Theta$ for $1 \leq i \leq m$, which together with axiom $\mathsf{A5}$, yields $\boxed{\mathsf{S}}\psi_i \in \Theta$ for $1 \leq i \leq m$. Hence, $\neg\langle\mathsf{S}\rangle\neg\mathsf{P} \in \Theta$. However, since Θ is an LM-MCS we have $\langle\mathsf{S}\rangle\mathsf{P} \to \langle\mathsf{S}\rangle\neg\mathsf{P} \in \Theta$, and thus $\langle\mathsf{S}\rangle\neg\mathsf{P} \in \Theta$. Contradiction. Consequently, Ω' is an LM-CS. Let Ω be the LM-MCS extending Ω' (Lindenbaum's lemma). By construction of $\mathfrak{M}^c$ we obtain $\Omega \in \mathcal{R}^{c}_{\boxed{\mathsf{S}}}(\Theta)$ and since $\neg\mathsf{P} \in \Omega' \subseteq \Omega$ we have $\Omega \notin W^{c}_{\mathsf{P}}$. By the assumption $\Theta \in W^{\Sigma}$ and the fact $\{\boxed{\mathsf{U}}\psi \mid \boxed{\mathsf{U}}\psi \in \Sigma\} \subseteq \{\boxed{\mathsf{U}}\psi \mid \boxed{\mathsf{U}}\psi \in \Theta\}$ (due to $\boxed{\mathsf{U}}\psi \to \boxed{\mathsf{U}}\boxed{\mathsf{U}}\psi \in \Sigma$) we know $\Omega \in W^{\Sigma}$, $\Omega \in \mathcal{R}^{c}_{\boxed{\mathsf{S}}}(\Theta) \cap W^{\Sigma} = \mathcal{R}^{\Sigma}_{\boxed{\mathsf{S}}}(\Theta)$, and $\Omega \notin W^{\Sigma}_{\mathsf{P}} = W^{c}_{\mathsf{P}} \cap W^{\Sigma}$.

R5 Similar to **R4**. QED

Theorem 5.2 (Strong Completeness of LM). *For any formula $\varphi \in \mathcal{L}^{\mathsf{LM}}$, and any $\Theta \subseteq \mathcal{L}^{\mathsf{LM}}$: if $\Theta \models_{\mathsf{LM}} \varphi$, then $\Theta \vdash_{\mathsf{LM}} \varphi$.*

Proof. The proof is by contraposition. Suppose φ is not LM-derivable from Θ. This means that $\Theta \cup \{\neg\varphi\}$ is a LM-CS. Namely, if $\Theta \cup \{\neg\varphi\}$ would be LM-inconsistent, then $\Theta, \neg\varphi \vdash_{\mathsf{LM}} \bot$ and so $\Theta \vdash_{\mathsf{LM}} \varphi$. By Lindenbaum's Lemma there is a $\Theta' \subseteq \mathcal{L}^{\mathsf{LM}}$ such that Θ' is a LM-MCS and $\Theta \cup \{\neg\varphi\} \subseteq \Theta'$. By the construction of the canonical model, $\Theta' \in W^c$ of $\mathfrak{M}^c$. By the truth lemma (cf. Lemma 4.5) we know that $\mathfrak{M}^c, \Theta' \models \Theta$ and $\mathfrak{M}^c, \Theta' \models \neg\varphi$. Let $\mathfrak{M}^{\Theta'}$ be a submodel of $\mathfrak{M}^c$ according to Definition 5.9. By definition $\Theta' \in W^{\Theta'}$ of $\mathfrak{M}^{\Theta'}$. By Lemma 5.7 we know $\mathfrak{M}^{\Theta'}, \Theta' \models \Theta$ and $\mathfrak{M}^{\Theta'}, \Theta' \models \neg\varphi$. Last, by Lemma 5.8, $\mathfrak{M}^{\Theta'}$ is an LM-model and so $\Theta \not\models_{\mathsf{LM}} \varphi$. QED

5.4 A Formal Analysis of Maṇḍana's Reduction

Here, we address Objective 2 and show that LM can adequately represent Maṇḍana's deontic theory. We provide a formal analysis of Maṇḍana's reduction and discuss whether the validity of four central Mīmāṃsā principles concerning commands is preserved through Maṇḍana's reduction.

At the heart of Maṇḍana's deontic theory lies the reduction of all deontic modalities to a uniform notion of instrumentality. Following Maṇḍana, our formal definition of instrumentality must satisfy the following criteria: (i) The instrument relation contains three components: an action Δ, serving as the instrument; a state of affairs φ, representing the outcome of Δ; and a state of affairs χ defining the circumstances in which Δ functions as an instrument for bringing about φ. (ii) The circumstances χ must be meaningful, which in Mīmāṃsā terms means that χ must be possible in the broadest sense (cf. not logically impossible).[10] Moreover, the agent in question must have a proper *choice* to execute action Δ when the appropriate circumstances χ occur. We split choice into a positive and negative component: (iii) Δ can be performed by the agent, and (iv) the agent can refrain from performing Δ. For a motivation of (i)–(iv), see Śabara on PMS 6.1 in (Subbāśāstrī, 1929-1934).[11]

We propose the defined instrumentality operator $\mathcal{I}(\Delta/\varphi/\chi)$ which is interpreted as follows:

> Δ is an instrument for guaranteeing φ in circumstances χ iff (i) If circumstance χ holds, performance of Δ guarantees φ, (ii) χ is possible, and if χ holds, both (iii) Δ is possible and (iv) $\overline{\Delta}$ is possible.

The corresponding formal definition, based on (i)-(iv), is given in Definition 5.10.

Based on the above, we can define Maṇḍana's reduction of the various command types to statements of instrumentality: obligatory and prohibited actions (denoted by $\mathcal{O}$, respectively $\mathcal{F}$) are defined in terms of those actions being instrumental to the reduction of bad karma (denoted by $\mathtt{R}$), and the accumulation of bad karma (denoted by $\mathtt{P}$), respectively. Elective commands (denoted by $\mathcal{E}$) are actions instrumental to outcomes that are neither $\mathtt{P}$ nor $\mathtt{R}$. Additionally, we need to ensure that the following Maṇḍana principle, which applies to obligations and prohibitions, is satisfied: "an action Δ cannot be an instrument for both the reduction $\mathtt{R}$ and the increase $\mathtt{P}$ of bad karma" (cf. Remark 5.1). This is done by introducing an additional clause requiring that the action in question is not simultaneously instrumental to the accumulation, respectively reduction of bad karma. We thus have that an action is obligatory (prohibited) if and only if it is an instrument for reducing (accumulating) bad karma *and* at the same time the action is not an instrument for accumulating (reducing) bad karma.

Definition 5.10. *Maṇḍana's notion of instrumentality is defined as:*

[10] We take the term 'meaningful' in this context to denote possibility in a general sense. To illustrate, the Mīmāṃsā consider the Vedic statement "one should build an altar in the sky" meaningless as a command, not because it is impossible to build an altar in the sky at this particular moment, but because it is conceptually impossible. For this reason, the Mīmāṃsā take this Vedic statement to express praise instead of a command. We use $\mathrlap{\,\scriptscriptstyle\mathsf{U}}\Diamond$ further below to formally represent this notion of possibility.

[11] See Chapter 3 and 4 for a discussion of the closely related notions of deontic contingency, respectively, deliberative agency.

$$
\begin{array}{lclll}
\mathcal{I}(\Delta/\varphi/\chi) & := & (i) & \boxed{\scriptstyle\mathsf{U}}(\chi \to \boxed{\scriptstyle\mathsf{S}}(t(\Delta) \to \varphi)) & \wedge \\
 & & (ii) & \Diamond_{\mathsf{U}}\chi & \wedge \\
 & & (iii) & \boxed{\scriptstyle\mathsf{U}}(\chi \to \Diamond_{\mathsf{S}} t(\Delta)) & \wedge \\
 & & (iv) & \boxed{\scriptstyle\mathsf{U}}(\chi \to \Diamond_{\mathsf{S}} \neg t(\Delta)) &
\end{array}
$$

Maṇḍana's reduction of obligations, prohibitions, and elective duties is defined as:

$$
\begin{array}{lcl}
\mathcal{O}(\Delta/\chi) & := & \mathcal{I}(\Delta/\mathtt{R}/\chi) \wedge \neg\mathcal{I}(\Delta/\mathtt{P}/\chi) \\[1em]
\mathcal{F}(\Delta/\chi) & := & \mathcal{I}(\Delta/\mathtt{P}/\chi) \wedge \neg\mathcal{I}(\Delta/\mathtt{R}/\chi) \\[1em]
\mathcal{E}(\Delta/\varphi/\chi) & := & \mathcal{I}(\Delta/\varphi/\chi) \textit{ with } \varphi \nvdash_{\mathsf{LM}} \mathtt{P} \textit{ and } \varphi \nvdash_{\mathsf{LM}} \mathtt{R}
\end{array}
$$

In the above definition of $\mathcal{I}(\Delta/\varphi/\chi)$, Δ refers to an action, φ to the outcome of that action (if applicable), and χ to the circumstances in which the instrumentality relation holds. The side condition on the elective duty $\mathcal{E}(./././)$ in Definition 5.10 ensures that results explicitly described by the command do not directly entail the accumulation or reduction of bad karma. However, indirectly this is allowed. We will see this when analyzing the Śyena controversy in Section 5.5. Last, we point out that obligations $\mathcal{O}(\Delta/\chi)$ could be equivalently defined as $\mathcal{I}(\Delta/\mathtt{R}/\chi) \wedge \neg\boxed{\scriptstyle\mathsf{U}}(\chi \to \boxed{\scriptstyle\mathsf{S}}(t(\Delta) \to \mathtt{P})$ due to the overlapping clauses (ii)-(iv) of the definition of instruments in $\mathcal{I}(\Delta/\mathtt{R}/\chi)$ and $\mathcal{I}(\Delta/\mathtt{P}/\chi)$. This also holds for prohibitions. The above definitions of the three command types ensure that Vedic actions can never be instrumental to both the reduction and the accumulation of bad karma (with electives leading to neither). This property is motivated by the Mīmāṃsā principle, endorsed by Maṇḍana, stating that "an action cannot be an instrument for both the reduction and the increase of bad karma" (Viraraghavacharya, 1971, on PMS 1.1.2).

Remark 5.2. *In* LM*, we define commands as having states of affairs as their condition. In addition, due to the translation* t *from the action language* $\mathcal{L}^{\mathsf{LAN}}$ *to the object level language* $\mathcal{L}^{\mathsf{LM}}$*, we can express prescriptions such as "offer to Agni once you have offered to Soma", which have as a condition an action that temporally precedes the prescribed action. This sentence formally corresponds to* $\mathcal{O}(\mathsf{Agni}/t(\mathsf{Soma}))$*, where* $t(\mathsf{Soma})$ *is the state of affairs witnessing that "the Soma offering has just been performed".*

We now show that important Mīmāṃsā properties hold for the derived deontic operators and that the Mīmāṃsā principles adopted by Maṇḍana are LM-valid formulae.

Irreducibility. Recall that for Mīmāṃsā authors, obligations, prohibitions, and elective duties are reciprocally irreducible (Freschi and Pascucci, 2021; Lellmann et al., 2021). Maṇḍana also adopts this view by limiting the type of results of the instruments corresponding to the three command types. That Definition 5.10 preserves this property is due to the second conjunct of the definitions of obligations and prohibitions and the side condition imposed on elective duties.

Contingency. For Mīmāṃsā, actions occurring in Vedic commands must be meaningful (cf. ŚBh on PMS 6.1, (Subbāśāstrī, 1929-1934)). An action is meaningful when an agent can perform the action and refrain from performing it. The property of meaningfulness of actions is expressed via the following LM-valid formula, which is a consequence of clauses (iii) and (iv) of Definition 5.10:

$$\mathcal{I}(\Delta/\varphi/\chi) \to \boxed{\mathsf{U}}(\chi \to (\Diamond_{\mathsf{S}} t(\Delta) \wedge \Diamond_{\mathsf{S}} \neg t(\Delta)))$$

where either $\varphi \in \{\mathtt{P}, \mathtt{R}\}$ or both $\varphi \nvdash_{\mathsf{LM}} \mathtt{P}$ and $\varphi \nvdash_{\mathsf{LM}} \mathtt{R}$. That is, the above holds for all three command types. In deontic logic, this property is known as the *contingency principle* (Anderson and Moore, 1957; von Wright, 1951) (See Chapter 3 for an extensive discussion). We point out that for obligations and prohibitions, the property is already implied by axioms A6 and A7, which ensure that the accumulation, respectively reduction, of bad karma can always be avoided. Consequently, in the light of these axioms, condition (iv) of instruments (Definition 5.10) is admissible for obligations and prohibitions but remains necessary for ensuring the meaningfulness of actions involved in elective commands.

No Impossible Commands. The logic LM implies the validity of a deontic consistency axiom for prohibitions (cf. the D-axiom of Standard Deontic Logic on page 13):

$$\neg(\mathcal{F}(\Delta/\chi) \wedge \mathcal{F}(\overline{\Delta}/\chi))$$

This valid formula corresponds to the Mīmāṃsā principle: "It is impossible that the Vedas tell you that you'll fall (i.e., be reborn in hell) both if you do something and if you don't do it" (Viraraghavacharya, 1971, p. 32). The quote illustrates the impossibility of the Vedas to give contradictory commands. The formula is valid due to the definition of instrumentality together with axiom A6. We obtain a similar LM-valid formula expressing this property for obligations:

$$\neg(\mathcal{O}(\Delta/\chi) \wedge \mathcal{O}(\overline{\Delta}/\chi))$$

As desired, the property does not hold for elective duties. This follows from the fact that these duties lead to worldly results on which no additional Mīmāṃsā property is imposed (see Definition 5.4 and 5.10).

Furthermore, LM satisfies the Mīmāṃsā principle that obligations and prohibitions are mutually exclusive, namely, no action Δ can be both obligatory and prohibited. The following LM-valid formula expresses this:

$$\neg(\mathcal{O}(\Delta/\chi) \wedge \mathcal{F}(\Delta/\chi))$$

The property is guaranteed by Definition 5.10. How obligations and prohibitions are defined implies that, in Maṇḍana's language, Δ cannot simultaneously be an *instrument* for the reduction and accumulation of bad karma. Still, from a semantic perspective, LM allows for situations where we end up at a world where both $\mathtt{P}$ and $\mathtt{R}$ hold after executing some action Δ (cf. Remark 5.1). However, in those cases Definition 5.10 ensures that this action Δ is neither obligatory nor prohibited. We refer to Chapter 3 for a discussion of the deontic consistency principle in the field of deontic logic.

Four General Mīmāṃsā Principles. The Mīmāṃsā are known for using *nyāya*s, i.e., general reasoning principles about commands. Three such principles were initially introduced and discussed by Ciabattoni et al. (2015), Freschi et al. (2017), and Lellmann et al. (2021). A fourth one was later discovered and introduced in (van Berkel et al., 2022a). We refer to these works for a detailed discussion of these principles. The logics of the Mīmāṃsā authors Prabhākara and Kumārila developed and discussed in (van Berkel et al., 2022a) are built upon these four principles. Maṇḍana conceptually deviates from the Mīmāṃsā tradition and does not take commands as primitive notions. Therefore, we investigate whether the validity of these principles is preserved by Maṇḍana's deontic reduction.

First, we briefly elaborate on these *nyāya*-based principles:

P1 If the accomplishment of an action presupposes the accomplishment of another action, the obligation to perform the first action implies the obligation to perform the second one. Conversely, if an action necessarily implies another prohibited action, the former is also prohibited.

P2 Two actions that exclude each other can neither be prescribed to nor prohibited for the same group of eligible people under the same conditions.

P3 If two sets of conditions identify the same group of eligible agents, then a command which holds for the conditions in one of those sets also holds for the conditions in the other set.

P4 If two fixed obligations are compatible, their joint performance is obligatory too.

Principle **P1**, constitutes the abstraction and reformulation of various *nyāya*s; among them, a *nyāya* present in the *Tantrarahasya* (IV.4.3.3) composed by the Mīmāṃsā author Rāmānujācārya (possibly 15th c. CE). The literal translation of this *nyāya* is "When the various [requirements of a given duty], beginning with the origination [of a new duty], are not established by other distinct prescriptions, then [the only prescription available] itself creates the other four prescriptions that are related to it". This quote signifies that if an obligatory action consists of sub-actions, these constitutive actions are also obligatory. Conversely, if an action is forbidden, all the composed rituals which include that action are forbidden too. To illustrate, given a prohibition to cross the ocean, working at a place across the ocean is also prohibited since it consists of (and requires) crossing the ocean.

Principle **P2** constitutes the abstract formulation of the so-called *principle of the half-hen* (see Kumārila's *Tantravārtika* ad 1.3.3 (Subbāśāstrī, 1929-1934)). In its most general form, this principle says that the collection of all Vedic commands is consistent, i.e., the performance (non-performance) of an obligatory (forbidden) action cannot lead to violating another Vedic command. Consequently, if some action is obligatory (forbidden), then neither is that action prohibited (obligatory) nor is its opposite obligatory (forbidden). Furthermore, **P2** strongly relates to the previously discussed Mīmāṃsā principle, according to which nothing impossible can be commanded (see page 189).

Principle **P3** comes from a discussion on the eligibility to perform sacrifices in ŚBh on PMS 6.1.25. It expresses the generality of prescriptions concerning extensionally and logically equivalent conditions.

Last, **P4** corresponds to a restricted form of the logical property known as *aggregation* (cf. Section 3.5). In Sanskrit, this property is called 'accumulation', i.e., *samuccaya*. In common Mīmāṃsā, cases of different fixed obligations to be performed in the same context are handled as follows: if the two actions are compatible with one another and are functional towards different intermediate results (e.g., brush your teeth and floss them, achieving different intermediate results even though both having the overall purpose of having healthy teeth), then one is obliged to perform them both. Otherwise, only one of the two must be performed, chosen according to various criteria. The *samuccaya* principle does not apply to elective sacrifices because even if the two were compatible, they would have the same purpose and, therefore, it is enough only to perform one of the two.[12] *Samuccaya* is defined for prescriptions and their results and does not apply to prohibitions.

The four Mīmāṃsā principles are phrased in terms of commands. To address whether they hold in LM, we have to reformulate them in terms of instrumentality statements. We adopt $\boxed{\mathsf{U}}$ to denote the relevant facts (e.g., the identification the same group of eligible agents in **P3**) under which conclusions can be drawn about Vedic commands. We now treat each principle in turn.

Principle **P1** consists of two formalizations, one for obligations and one for prohibitions:

p1a $(\mathcal{O}(\Delta/\chi) \wedge \boxed{\mathsf{U}}(t(\Delta) \rightarrow t(\Gamma))) \rightarrow \mathcal{O}(\Gamma/\chi)$

p1b $(\mathcal{F}(\Delta/\chi) \wedge \boxed{\mathsf{U}}(t(\Gamma) \rightarrow t(\Delta))) \rightarrow \mathcal{F}(\Gamma/\chi)$

p1a and **p1b** are not LM-valid formulae (it is straightforward to construct a countermodel). This is desirable: instrumentality is a notion of *sufficient* means, not of necessary means such as expressed in **P1**. Maṇḍana appears to be aware of the fact that **P1** does not hold in his theory. To preserve the gist of **P1**, Maṇḍana provides an explicit explanation of the role of necessary preconditions as independent from instrumentality. Namely, Maṇḍana takes the reduction of bad karma as a universally desired goal. Then, he argues, from a rational point of view no agent would be willing to omit those actions $\Gamma_1, \ldots, \Gamma_n$ that

[12] *tayor ekārthatvāt samuccayo na sambhavati* (ŚBh 8.1.15.26), "Since the two [actions] have the same purpose, aggregation (*samuccaya*) is impossible".

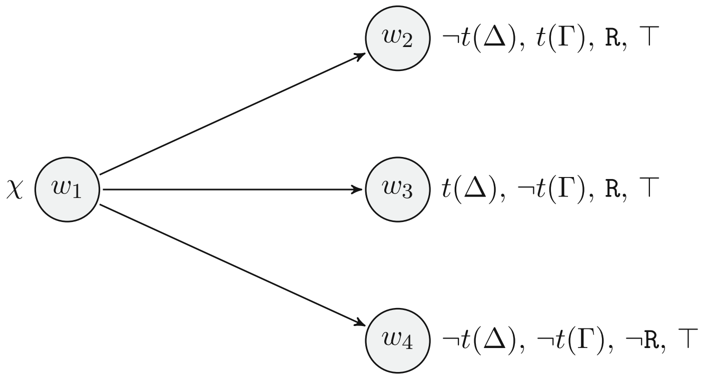

w_2
$\neg t(\Delta)$, $t(\Gamma)$, $\mathtt{R}$, $\top$
χ
w_1
w_3
$t(\Delta)$, $\neg t(\Gamma)$, $\mathtt{R}$, $\top$
w_4
$\neg t(\Delta)$, $\neg t(\Gamma)$, $\neg \mathtt{R}$, $\top$

Figure 5.1: Counterexample showing that principle **p2** is not LM-valid for obligations and elective duties, $\varphi = \mathtt{R}$ for the obligation reading of **p2** and $\varphi = \top$ for the elective duty reading of **p2**. The $\mathcal{R}_{\boxed{\mathsf{U}}}$-relation is left implicit, the arrows refer to $\mathcal{R}_{\boxed{\mathsf{S}}}$.

serve as necessary preconditions for an obligatory action Δ which reduces bad karma, even though the actions $\Gamma_1, \ldots, \Gamma_n$ may not be sufficient (i.e., instruments) for reducing bad karma.

We point out that an alternative formalization of **P1** is possible where the factual condition in the antecedent of **p1a** and **p1b** is relativized to the circumstance χ, namely, $\boxed{\mathsf{U}}(\chi \to \boxed{\mathsf{S}}(t(\Delta) \to t(\Gamma)))$, respectively, $\boxed{\mathsf{U}}(\chi \to \boxed{\mathsf{S}}(t(\Gamma) \to t(\Delta)))$. For the same reasons given for **p1a** and **p1b**, the resulting formalizations are not LM-valid formulae.

p2 $(\mathcal{I}(\Delta/\varphi/\chi) \wedge \boxed{\mathsf{U}}(t(\Delta) \to \neg t(\Gamma))) \to \neg \mathcal{I}(\Gamma/\varphi/\chi)$ such that (†) holds

(†) it is the case that either $\varphi \in \{\mathtt{P}, \mathtt{R}\}$ or both $\varphi \nvdash_{\mathsf{LM}} \mathtt{P}$ and $\varphi \nvdash_{\mathsf{LM}} \mathtt{R}$.

Condition (†) ensures that the three properties are defined for obligations, prohibitions, and elective duties. To illustrate, **p2** as defined for obligations can be read as $(\mathcal{O}(\Delta/\chi) \wedge \boxed{\mathsf{U}}(t(\Delta) \to \neg t(\Gamma))) \to \neg \mathcal{O}(\Gamma/\chi)$. Principle **p2** is not LM-valid.[13] The counterexample in Figure 5.1 shows this for obligations and elective duties (a similar countermodel can be straightforwardly constructed for prohibitions). We make three observations: First, actions Δ and Γ can be two mutually exclusive instruments serving the same purpose; for instance, in the case of elective duties, they may both lead to a tautological outcome $\top$. Second, Δ and Γ cannot be jointly performed. However, that these actions are mutually exclusive does not mean that not performing one of the two actions leads to a deontically undesirable outcome. In fact, in the case of obligations, one of Δ and Γ suffices for

[13] In (van Berkel et al., 2022a), principle **P2** was formalized for conflicting outcomes instead of actions, i.e., $(\mathcal{I}(\Delta/\varphi/\chi) \wedge \boxed{\mathsf{U}}(\varphi \to \psi)) \to \neg \mathcal{I}(\Delta/\psi/\chi)$. This formula is LM-valid. The logics in the aforementioned work—formalizing the deontic theories of Prabhākara and Kumārila—do not distinguish between obligatory actions and obligatory outcomes. However, in LM, we can distinguish between the two. Therefore, the formalization expressed in **p2** is a more accurate rendition of **P2**.

ascertaining a reduction of bad karma (cf. the heuristics applied for **P4** discussed in Section 5.1). Third, in cases such as the one described above, the Mīmāṃsā principle called *Vikalpa* applies (Freschi and Pascucci, 2021). The principle states that in case two (or more) commands conflict, and no other heuristics can be applied to resolve the conflict, one can choose to follow *one* of these commands (Lellmann et al., 2021). This universally accepted Mīmāṃsā principle only serves as a last resort in case of conflict. Nevertheless, *Vikalpa* requires that the agent complies with at least one of the conflicting commands. There is a strong connection between *Vikalpa* and the nonmonotonicity principle known as *disjunctive response* (Horty, 2012) (see Section 3.5).

Here too, **P2** can be alternatively formalized relativizing the factual condition in the antecedent of **p2** to the circumstance χ, namely, $\boxed{\scriptstyle \mathsf{U}}(\chi \rightarrow \boxed{\scriptstyle \mathsf{S}}(t(\Delta) \rightarrow \neg t(\Gamma)))$. The model given in Figure 5.1 also serves as a countermodel to this interpretation.

p3 $(\mathcal{I}(\Delta/\varphi/\chi) \wedge \boxed{\scriptstyle \mathsf{U}}(\chi' \equiv \chi)) \rightarrow \mathcal{I}(\Delta/\varphi/\chi')$ such that (†) holds

(†) it is the case that either $\varphi \in \{\mathtt{P}, \mathtt{R}\}$ or both $\varphi \nvdash_{\mathsf{LM}} \mathtt{P}$ and $\varphi \nvdash_{\mathsf{LM}} \mathtt{R}$.

The formalization in **p3** is defined to hold for all three command types. Principle **p3** is LM-valid, namely, because the universal necessity modality $\boxed{\scriptstyle \mathsf{U}}$ is a normal modal operator. The principle captures a minimal property of logical reasoning: its absence would make a formalized prescription dependent upon the particular form of a formula, e.g., a circumstance χ would be different from the circumstance $\chi \wedge \chi$. Since the factual condition in **p3**—i.e., $\boxed{\scriptstyle \mathsf{U}}(\chi' \equiv \chi)$—is already about the relevant circumstances, we do not obtain an alternative formalization of **p3** as we did for **p1a**, **p1b**, and **p2**.

Last, principle **P4** expresses a restricted form of aggregation for compatible actions. Intuitively, two actions $t(\Delta)$ and $t(\Gamma)$ are compatible if their joint performance is possible. There are two ways in which we can then formalize this idea of possibility: **p4a** expresses the interpretation of 'compatible' as a global notion—i.e., $\langle\mathsf{U}\rangle(t(\Delta) \wedge t(\Gamma))$—whereas **p4b** interprets 'compatible' locally, that is, relativized to the relevant circumstances χ—i.e., $\boxed{\scriptstyle \mathsf{U}}(\chi \rightarrow \langle\mathsf{S}\rangle(t(\Delta) \wedge t(\Gamma)))$. Both **p4a** and **p4b** are formulated for instruments in general (recall that & denotes joint action, page 177).

p4a $(\langle\mathsf{U}\rangle(t(\Delta) \wedge t(\Gamma)) \wedge \mathcal{I}(\Delta/\varphi/\chi) \wedge \mathcal{I}(\Gamma/\psi/\chi)) \rightarrow \mathcal{I}(\Delta\&\Gamma/\varphi \wedge \psi/\chi)$

p4b $(\boxed{\scriptstyle \mathsf{U}}(\chi \rightarrow \langle\mathsf{S}\rangle(t(\Delta) \wedge t(\Gamma))) \wedge \mathcal{I}(\Delta/\varphi/\chi) \wedge \mathcal{I}(\Gamma/\psi/\chi)) \rightarrow \mathcal{I}(\Delta\&\Gamma/\varphi \wedge \psi/\chi)$

Formula **p4a** is not an LM-valid formula, whereas the local interpretation of **P4**, expressed by **p4b**, is in fact LM-valid. An intuitive explanation for this is the following: since φ, respectively ψ, is propagated at all worlds that witness performances of Δ, respectively Γ, both φ and ψ are also propagated at those worlds witnessing $\Delta\&\Gamma$. The fact that $\Delta\&\Gamma$ is possible, together with the fact that $\langle\mathsf{S}\rangle\neg t(\Delta)$ implies $\langle\mathsf{S}\rangle\neg t(\Delta\&\Gamma)$, ensures that $\boxed{\scriptstyle \mathsf{U}}\chi \rightarrow$

$\Diamond_{\mathsf{S}}\neg t(\Delta\&\Gamma)$ holds, which is a necessary condition for instrumentality (Definition 5.10). Property **p4b** holds for instruments in general and thus also for commands defined in terms of them. For instance, we obtain the following LM-valid formulae for obligations:

$$(\boxed{\mathsf{U}}(\chi \to (\Diamond_{\mathsf{S}}(t(\Delta) \wedge t(\Gamma))) \wedge \mathcal{O}(\Delta/\chi) \wedge \mathcal{O}(\Gamma/\chi)) \to \mathcal{O}(\Delta\&\Gamma/\chi)$$

Recall that for the Mīmāṃsā, principle **P4** is only adopted for obligations. There is no such distinction in Maṇḍana's deontic theory because the three command types are just instruments whose possible joint performance gives rise to other instruments. On this general level, there is no difference between aggregating consistent obligations or prohibitions.[14]

5.5 The Śyena Controversy

The introduced logic LM enables us to understand better the underlying structure of Maṇḍana's deontic theory. Here, we apply LM to reconstruct and formally verify Maṇḍana's solution to the Śyena controversy. Our aim (Objective 3) is to show and explain the consistency of Maṇḍana's solution to the controversy. The Śyena is a one-day-long ritual in which the Soma beverage is offered. Its putative result is the death of the sacrificer's enemy. The controversy is due to the fact that the Śyena appears to be prescribed in the Vedas (through the command "The one who desires to kill their enemy should sacrifice bewitching with the Śyena"), even though the Vedas also prohibits harming any living being. Thus, performing Śyena seems to conflict with the prohibition to harm. However, the Vedas are unanimously assumed consist. Finding a solution to the controversy proved challenging for many Mīmāṃsā scholars. We refer to (van Berkel et al., 2022a) for an extensive discussion of various authors and their solutions to the Śyena controversy. The Śyena controversy consists of the following four statements:

(A) The one who desires to kill their enemy should sacrifice with the Śyena.

(B) One should not harm any living being.

(C) Performing Śyena implies causing someone's death.

(D) Causing someone's death implies harming.

We point out that (A) and (B) are direct translations from Sanskrit, whereas (C) and (D) are derived from common Mīmāṃsā arguments about the Śyena. Furthermore, all Mīmāṃsā authors agree that Śyena is an elective sacrifice (A) and that the command not to harm any living being (B) is a prohibition. Although all Mīmāṃsā authors agree that the Śyena should not be performed, they disagree on the reasons underlying it.

[14] See Section 3.5 and the work of Parent and van der Torre (2018b) for a discussion on restricted forms of aggregation.

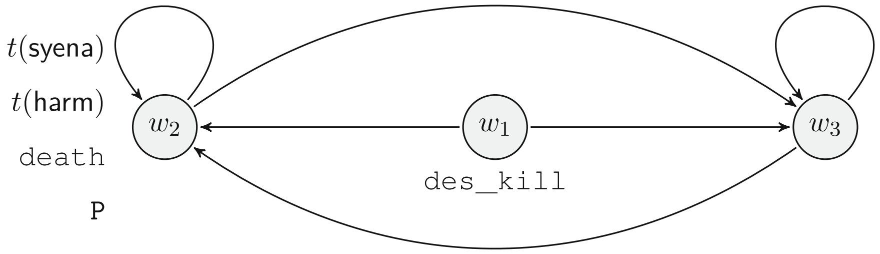


Figure 5.2: The Śyena model $\mathfrak{M} = \langle \mathfrak{F}, V \rangle$ with $\mathsf{Atoms} = \{\texttt{death}, \texttt{des\_kill}\}$ and $\mathsf{Wit} = \{\mathsf{syena}, \mathsf{harm}\}$, $W = \{w_1, w_2, w_3\}$, $W_{\mathsf{syena}} = W_{\mathsf{harm}} = W_{\texttt{P}} = \{w_2\}$, $W_{\texttt{R}} = \emptyset$, $\mathcal{R}_{\boxed{\mathsf{S}}} = \{(w_1, w_2), (w_1, w_3), (w_2, w_2), (w_2, w_3), (w_3, w_2), (w_3, w_3)\}, V(\texttt{des\_kill}) = \{w_1\}$, and $V(\texttt{death}) = \{w_2\}$. Arrows represent the relation $\mathcal{R}_{\boxed{\mathsf{S}}}$. The relation $\mathcal{R}_{\boxed{\mathsf{U}}} = W \times W$ is left implicit. The sentences $(\mathsf{A}_{\mathsf{LM}})$–$(\mathsf{D}_{\mathsf{LM}})$ are in fact satisfied at each world.

In what follows, `we use this font` to denote states of affairs, and $\mathsf{we\ use\ this\ font}$ to denote actions. For the formalization of the controversy, let the action language $\mathcal{L}^{\mathsf{Act}}$ consist of the atoms syena and harm, respectively describing "performing the Śyena" and "doing harm". Let $\texttt{death}$ and $\texttt{des\_kill}$ be propositional atoms denoting "the death of the enemy has occurred" and "the agent has the desire to kill the enemy". The Śyena controversy is formalized in LM as follows:

$(\mathsf{A}_{\mathsf{LM}})$ $\mathcal{E}(\mathsf{syena}/\texttt{death}/\texttt{des\_kill}) = \mathcal{I}(\mathsf{syena}/\texttt{death}/\texttt{des\_kill})$

$(\mathsf{B}_{\mathsf{LM}})$ $\mathcal{F}(\mathsf{harm}/\top) = \mathcal{I}(\mathsf{harm}/\texttt{P}/\top) \wedge \neg\mathcal{I}(\mathsf{harm}/\texttt{R}/\top)$

$(\mathsf{C}_{\mathsf{LM}})$ $\boxed{\mathsf{U}}(t(\mathsf{syena}) \rightarrow \texttt{death})$

$(\mathsf{D}_{\mathsf{LM}})$ $\boxed{\mathsf{U}}(\texttt{death} \rightarrow t(\mathsf{harm}))$

The LM-model $\mathfrak{M} = \langle \mathfrak{F}, V \rangle$ shows the consistency of $(\mathsf{A}_{\mathsf{LM}})$-$(\mathsf{D}_{\mathsf{LM}})$ in LM, where:

- $\mathsf{Atoms} = \{\texttt{death}, \texttt{des\_kill}\}$ and $\mathsf{Wit} = \{\mathsf{syena}, \mathsf{harm}\}$;
- $W = \{w_1, w_2, w_3\}$, $W_{\mathsf{syena}} = W_{\mathsf{harm}} = W_{\texttt{P}} = \{w_2\}$, $W_{\texttt{R}} = \emptyset$;
- $\mathcal{R}_{\boxed{\mathsf{S}}} = \{(w_1, w_2), (w_1, w_3), (w_2, w_2), (w_2, w_3), (w_3, w_2), (w_3, w_3)\}$;
- $V(\texttt{des\_kill}) = \{w_1\}$ and $V(\texttt{death}) = \{w_2\}$.

Observe that, by Definition 5.10, if an instrumentality statement holds at some moment w of a model, then the instrumentality relation consisting of clauses (i)-(iv) of Definition 5.10 holds true at every moment of the model. Figure 5.2 represents the model $\mathfrak{M}$ graphically. It is the case that $\mathfrak{M} \models \boxed{\mathsf{U}}(\texttt{des\_kill} \rightarrow \boxed{\mathsf{S}}(t(\mathsf{syena} \rightarrow \texttt{death}))$, and

$\mathfrak{M} \models \Diamond_{\mathsf{U}}\, \mathtt{des\_kill}$. Furthermore, we have $\mathfrak{M} \models \Box_{\mathsf{U}}(\mathtt{des\_kill} \to \Diamond_{\mathsf{S}} t(\mathsf{syena}))$ and also $\mathfrak{M} \models \Box_{\mathsf{U}}(\mathtt{des\_kill} \to \Diamond_{\mathsf{S}} \neg t(\mathsf{syena}))$. So all the conditions (i)-(iv) of the instrumentality relation are satisfied. Consequently, we have that $\mathfrak{M} \models \mathcal{I}(\mathsf{syena}/\mathtt{death}/\mathtt{des\_kill})$ $(\mathsf{A}_{\mathsf{LM}})$. In a similar way we can verify that $\mathfrak{M}$ satisfies $(\mathsf{B}_{\mathsf{LM}})$, $(\mathsf{C}_{\mathsf{LM}})$ and $(\mathsf{D}_{\mathsf{LM}})$.

What is more, the sentences in $\Sigma = \{(\mathsf{A}_{\mathsf{LM}}), (\mathsf{C}_{\mathsf{LM}}), (\mathsf{D}_{\mathsf{LM}})\}$ logically entail that the Śyena is prohibited. Namely, we have that:

$$\Sigma \models_{\mathsf{LM}} \mathcal{F}(\mathsf{syena}/\mathtt{des\_kill})$$

The reasoning is straightforward. First, observe that A_{LM} entails the conditions (ii)-(iv) of the definition of $\mathcal{F}(\mathsf{syena}/\mathtt{des\_kill})$ (Definition 5.10). For condition (i), observe that condition (i) of $(\mathsf{B}_{\mathsf{LM}})$ expresses that harm necessarily leads to the accumulation of bad karma $\mathtt{P}$. Since syena necessarily leads to $\mathtt{death}$ $(\mathsf{C}_{\mathsf{LM}})$ and $\mathtt{death}$ necessarily entails harm $(\mathsf{D}_{\mathsf{LM}})$, the performance of syena necessarily leads to the accumulation of bad karma $\mathtt{P}$.

Recall the footnote on page 180. The conclusion $\mathcal{F}(\mathsf{syena}/\mathtt{des\_kill})$ is not a *Vedic* prohibition. That is, there is no such statement in the sacred texts. The prohibition $\mathcal{F}(\mathsf{syena}/\mathtt{des\_kill})$ follows on the *derived* level. On this level, it is possible that Maṇḍana's theory implies elective commands and prohibitions for the same actions without leading to an inconsistency.

For Maṇḍana, then, there is a dilemma. It is true that "if you desire to kill your enemy, you are commanded to sacrifice with the Śyena", but also "if you desire to kill your enemy, you are prohibited from performing the Śyena". He solves this dilemma not on a deontic level but by an appeal to rationality. Maṇḍana argues that the Śyena should not be performed because even though it provides the worldly reward of the death of one's enemy, this reward is strictly outweighed by the accumulation of bad karma, which is a necessary consequence of performing the Śyena. Maṇḍana distinguishes between two kinds of desires: worldly desires (such as desiring the death of one's enemy) and karma desires (the desire to diminish one's accumulated bad karma and the desire not to accumulate bad karma). According to Maṇḍana, the last kind of desire is necessary for any rational being. Whereas Maṇḍana does not distinguish between types of worldly desires, he does distinguish between worldly desires and karma desires. Our formalization mirrors this: karma desires–i.e., $\mathtt{P}$ and $\mathtt{R}$—define obligation and prohibition, and worldly desires define elective duties. Additionally, Maṇḍana differentiates between the strengths of worldly desires and those of karma desires. Namely, to resolve the apparent decision dilemma of the Śyena controversy, Maṇḍana appeals to the different strengths of the two desires involved: i.e., no rational agent would prefer worldly desires over karma-desires in case of conflict. The formal representation of such strengths, thus incorporating Maṇḍana's account of rational decision-making, is not present in the setting of LM. Such an extension is left for future work.

As a last remark, we point out that the Śyena controversy is based on actual prescriptions found in the Vedas. Maṇḍana interprets the controversy as a dilemma: on the one hand, there is the elective command to *perform* the Śyena and, on the other hand, there is a (derived) prohibition that forbids the performance of Śyena. For Maṇḍana, these commands are two incompatible instruments that hold in the same context. However, Maṇḍana's solution additionally appeals to the rationality of the agent, who would never perform the Śyena. As a benchmark challenge, the Śyena controversy fulfills a similar role for the Mīmāṃsā as deontic puzzles and paradoxes do in deontic logic.

5.6 Maṇḍana and the Deontic Paradoxes

In deontic logic, paradoxes and puzzles are the driving force for defining and refining formal systems. They serve as benchmarks and are quintessential for properties that (intuitively) should or should not hold in a deontic logic (see Chapter 1). This section investigates how the logic LM deals with well-known puzzles (Objective 4). We consider a selection of those paradoxes and refer to (van Berkel et al., 2022a) for an extensive analysis of a variety of paradoxes in relation to the logics based on Maṇḍana, Prabhākara, and Kumārila.

Remark 5.3. *Observe that for common Mīmāṃsā, negative obligations and prohibitions are two strictly distinct deontic concepts (Section 5.1). This is contrary to standard approaches in deontic logic where the two are taken as interdefinable (Hilpinen and McNamara, 2013). A notable exception in this respect is Talmudic logic (Abraham et al., 2011). Since the logic* LM *has the expressivity to differentiate between obligations and prohibitions, we consider several formal interpretations of the paradoxes containing both obligations and prohibitions. We omit consideration of elective duties (i.e., $\mathcal{E}$), which are identified through the presence of a desire.*

5.6.1 Contrary-to-Duty Paradoxes

Recall that contrary-to-duty (CTD) scenarios describe subideal situations in which obligations hold as a result of violating some other obligation (Chapter 1). In deontic dynamic logics—i.e., deontic PDL-like logics—obligations are about actions, and actions are taken as transitions between worlds (Meyer, 1988).[15] Most CTD scenarios are then straightforwardly addressed by assigning an inherently *temporal interpretation*. For instance, the CTD obligation "If you do not keep your promise, you ought to apologize" is interpreted as "*after* you do not keep your promise, you ought to apologize". The reason inconsistencies are avoided is that primary obligations are interpreted at a different moment from the secondary obligation, namely, after the occurrence of the violation. In the context of LM, we can consistently address these CTD paradoxes by adopting the same (temporal) approach taken by Meyer (1988). Due to the presence of action witnesses, we may formalize a CTD obligation, such as the one above, in the following two ways:

[15] Meyer (1988) discusses several paradoxes in the context of dynamic deontic logic.

- $[\overline{\mathsf{promise}}]\mathcal{O}(\mathsf{apologize}/\top)$
- $\mathcal{O}(\mathsf{apologize}/\neg t(\mathsf{promise}))$

The first is in the spirit of deontic dynamic logic (Meyer, 1988) (recall that the bar refers to an action's complement), and the second employs the notion of an action witness by conditioning the obligation on whether "not keeping one's promise" has been successfully performed at the present moment. Since temporal readings of CTD obligations have a straightforward solution, we will not pursue the above two options here. Moreover, as pointed out by Prakken and Sergot (1996), not all CTD scenarios can be resolved by introducing temporal distinctions. In this respect, we consider the Gentle Murder Paradox. Initially proposed by Forrester (1984), the paradox consists of the following four sentences:

(G1) You ought not to kill.

(G2) If you kill, you ought to kill gently.

(G3) Killing gently is killing.

(G4) You do kill.

In Standard Deontic Logic (Hilpinen and McNamara, 2013) (see page 13), the above sentences entail the following undesirable obligation:

(G5) You ought to kill.

First, we must rewrite the Gentle Murder Paradox in terms of the vocabulary of Maṇḍana, using instruments, rewards, and sanctions. A prohibition such as "you should not kill" can be interpreted as "the act of killing leads to the accumulation of bad karma", which is formalized as

$$\mathcal{F}(\mathsf{kill}/\top) := \mathcal{I}(\mathsf{kill}/\mathtt{P}/\top) \wedge \neg\mathcal{I}(\mathsf{kill}/\mathtt{R}/\top) \tag{1}$$

A sentence such as "if you kill, you ought to kill gently" can be read as "given that you kill, if you also kill gently, a reduction of bad karma ensues". In other words, the act of both killing and killing gently is instrumental to reducing bad karma. The sentence is formalized as,

$$\mathcal{O}(\mathsf{kill}\&\mathsf{gently}/\top) := \mathcal{I}(\mathsf{kill}\&\mathsf{gently}/\mathtt{R}/\top) \wedge \neg\mathcal{I}(\mathsf{kill}\&\mathsf{gently}/\mathtt{P}/\top) \tag{2}$$

At first, this formalization might seem odd. The above two formalizations suggest that the agent has a prohibition against killing and simultaneously an obligation to kill *and* do it gently. The scenario has the appearance of a dilemma. Should the agent kill or not? To better understand the situation, we must adopt Maṇḍana's perspective on the matter and think of the two formalizations as *instrumentality relations* instead. Vedic prohibitions and obligations are nothing but descriptions of instrumentality relations about the accumulation, respectively reduction of bad karma. This means that (1) and (2) express that killing is instrumental to accumulating bad karma, and killing gently is instrumental to reducing bad karma. If an agent desires to accumulate bad karma (which, according to Maṇḍana, a rational agent never will), the agent can kill. If an agent desires to reduce bad karma (which, according to Maṇḍana, a rational agent always will), the agent can employ the act of killing gently. In Maṇḍana's approach, a rational agent will never kill gently in order to reduce bad karma since killing gently inevitably leads to the abominable accumulation of bad karma (see page 196).

Since obligations and prohibitions are not interdefinable for the Mīmāṃsā, we obtain the following possible formalizations of this paradox:

(g1) $\mathcal{O}(\overline{\mathsf{kill}}/\top)$ or $\mathcal{F}(\mathsf{kill}/\top)$

(g2) $\mathcal{O}(\mathsf{kill\&gently}/\top)$ or $\mathcal{F}(\overline{\mathsf{gently}}\mathsf{\&kill}/\top)$

(g3) $\boxed{u}(t(\mathsf{gently}) \to t(\mathsf{kill}))$

(g4) $t(\mathsf{kill})$

Three of the four formalizations are consistent. Table 5.1 shows the different possibilities of formalizing sentences (g1) and (g2), and their consistency together with (g3) and (g4). Surprisingly, the only inconsistent option arises when (G1) is interpreted as $\mathcal{F}(\mathsf{kill}/\top)$ and (G2) as $\mathcal{O}(\mathsf{kill\&gently}/\top)$, i.e., (1) and (2) above. The inconsistency is explained as follows. By Definition 5.10 we know that $\mathcal{F}(\mathsf{kill}/\top)$ implies $\boxed{s}(t(\mathsf{kill}) \to \mathtt{P})$ and that $\mathcal{O}(\mathsf{kill\&gently}/\top)$ implies $\neg\boxed{s}((t(\mathsf{kill}) \wedge t(\mathsf{gently})) \to \mathtt{P})$ (recall that $\mathsf{kill\&gently} = \overline{\overline{\mathsf{kill}} \cup \overline{\mathsf{gently}}}$). With basic LM-reasoning we obtain $\Diamond_{s}(t(\mathsf{kill}) \wedge t(\mathsf{gently}) \wedge \neg\mathtt{P})$, which is inconsistent with the fact that $\boxed{s}(t(\mathsf{kill}) \to \mathtt{P})$. In other words, it conflicts with the Mīmāṃsā property that actions such as $\mathsf{kill\&gently}$ cannot be obligatory and prohibited simultaneously. We point out that, in general, an action may imply both $\mathtt{P}$ and $\mathtt{R}$ (cf. Remark 5.1). However, that action can neither be obligatory nor prohibited by Definition 5.10. The only correct Maṇḍana-like interpretation of the paradox is either the one in the first, third, or fourth row.

Consider the formal interpretation in the first row of Table 5.1. This interpretation is consistent. It expresses that not killing leads to a reduction of bad karma and that gentle killing leads to a reduction of bad karma. However, it does not entail that killing itself leads to a similar reduction. This interpretation entails the following LM-valid formulae

$$(3) \quad \mathcal{O}(\overline{\mathsf{kill}}/\top) \wedge \mathcal{O}(\mathsf{kill\&gently}/\top) \to \mathcal{O}(\overline{\mathsf{kill}} \cup (\mathsf{kill\&gently})/\top)$$

and

$$(4) \quad \mathcal{O}(\overline{\mathsf{kill}}/\top) \wedge \mathcal{O}(\mathsf{kill\&gently}/\top) \to \neg\mathcal{O}(\mathsf{kill}/\top)$$

In the above, (3) expresses that the paradox entails either not killing or killing gently suffices (as an instrument) to reduce bad karma. Furthermore, (4) expresses that although

You ought not to kill	If you kill, you ought to kill gently	Killing gently is killing	You kill	Consistent
$\mathcal{O}(\overline{\mathsf{kill}}/\top)$	$\mathcal{O}(\mathsf{kill}\&\mathsf{gently}/\top)$	$\boxed{\scriptstyle\mathsf{U}}(t(\mathsf{gently}) \to t(\mathsf{kill}))$	$t(\mathsf{kill})$	yes
$\mathcal{F}(\mathsf{kill}/\top)$	$\mathcal{O}(\mathsf{kill}\&\mathsf{gently}/\top)$	$\boxed{\scriptstyle\mathsf{U}}(t(\mathsf{gently}) \to t(\mathsf{kill}))$	$t(\mathsf{kill})$	no
$\mathcal{O}(\overline{\mathsf{kill}}/\top)$	$\mathcal{F}(\overline{\mathsf{gently}}\&\mathsf{kill}/\top)$	$\boxed{\scriptstyle\mathsf{U}}(t(\mathsf{gently}) \to t(\mathsf{kill}))$	$t(\mathsf{kill})$	yes
$\mathcal{F}(\mathsf{kill}/\top)$	$\mathcal{F}(\overline{\mathsf{gently}}\&\mathsf{kill}/\top)$	$\boxed{\scriptstyle\mathsf{U}}(t(\mathsf{gently}) \to t(\mathsf{kill}))$	$t(\mathsf{kill})$	yes

Table 5.1: The four possible formalizations of the Gentle Murder Paradox.

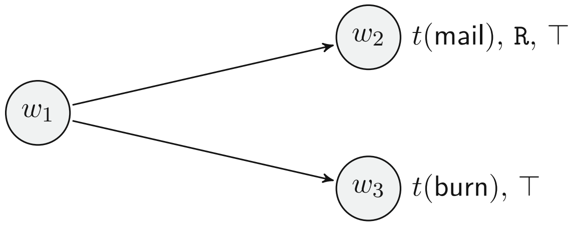

w_1
w_2
$t(\mathsf{mail})$, $\mathtt{R}$, $\top$
w_3
$t(\mathsf{burn})$, $\top$

Figure 5.3: Counterexample showing that Ross' Paradox does not hold in LM for the case of obligations, i.e., $w_1 \models \mathcal{O}(\mathsf{mail}/\top)$ and $w_1 \not\models \mathcal{O}(\mathsf{mail} \cup \mathsf{burn}/\top)$. The $\mathcal{R}_{\boxed{\scriptstyle\mathsf{U}}}$ relation is left implicit; the arrows refer to $\mathcal{R}_{\boxed{\scriptstyle\mathsf{S}}}$. The model is defined accordingly: $\mathfrak{M}_{\mathsf{M}} = \langle \mathcal{F}_{\mathsf{M}}, V \rangle$, with $W = \{w_1, w_2, w_3\}$, $W_{\mathsf{mail}} = \{w_2\}$, $W_{\mathsf{burn}} = \{w_3\}$, $W_{\mathsf{P}} = \emptyset$, $W_{\mathtt{R}} = \{w_2\}$, $\mathcal{R}_{\boxed{\scriptstyle\mathsf{U}}} = W \times W$, $\mathcal{R}_{\boxed{\scriptstyle\mathsf{S}}} = \{(w_1, w_2), (w_1, w_3)\}$ and V as defined in Definition 5.5.

an agent is obliged to kill gently from the point of view of instrumentality, the agent does not have an obligation to kill in general, as desired.

Last, we point out that the interpretations $\mathcal{F}(\mathsf{kill}/\top)$ and $\mathcal{F}(\overline{\mathsf{gently}}\&\mathsf{kill}/\top)$ express a more intuitive reading of (G1), respectively (G2) in terms of instruments. Namely, it tells us that killing is instrumental to an accumulation of bad karma and non-gentle killing is not an exception.

5.6.2 Ross' Paradox

One of the oldest deontic paradoxes is Ross' Paradox. Initially proposed by Ross (1944), the paradox consists of deriving from the obligation in (R1) the counterintuitive obligation expressed in (R2).

(R1) It is obligatory that agent i mails the letter.

(R2) It is obligatory that agent i mails the letter or burns it.

In most normal modal deontic logics, such as Standard Deontic Logic (page 13), the sentence (R2) is a logical consequence of (R1).[16] Ross' Paradox does not hold in LM. First, observe that the sentence expressed in (R1) can be formalized as either $\mathcal{O}(\mathsf{mail}/\top)$ or $\mathcal{F}(\neg\mathsf{mail}/\top)$. We consider two formal interpretations: one in which both sentences are taken as obligations and one in which both are considered as prohibitions. The countermodel in Figure 5.3 shows that the following formula is not LM-valid:

$$(5)\quad \mathcal{O}(\mathsf{mail}/\top) \rightarrow \mathcal{O}(\mathsf{mail} \cup \mathsf{burn}/\top)$$

In other words, Ross' paradox does not hold. A similar countermodel for the prohibition interpretation of the Paradox can be straightforwardly obtained. The reason why (5) is not LM-valid is that condition (i) of Definition 5.10 is not satisfied: only mailing the letter suffices for the reduction of bad karma. To understand this point, recall the discussion of principle **p1** in Section 5.4. Instrumentality is a notion of sufficient means, not of necessary means. In the case of obligations, even if the action witness $t(\mathsf{mail})$ implies the action $t(\mathsf{mail} \cup \mathsf{burn})$ (straightforward LM reasoning), the fact that mailing the letter is an instrument for the reduction of bad karma does not mean that either mailing or burning the letter is an instrument for that same purpose. The solution is similar to the one provided for dynamic deontic logics, e.g., by Meyer (1988).

5.6.3 The Alternative Service Challenge

The Alternative Service Challenge was extensively discussed by Horty (1994) in the context of nonmonotonic deontic reasoning. One of the earliest versions of the paradox can be found in the work of Van Fraassen (1973). It consists of the following three sentences:

(A1) You should fight in the army or perform alternative service.

(A2) You should not fight in the army.

(A3) You should perform alternative service.

Ideally, the sentence (A3) is logically implied by (A1) and (A2) together. The paradox is especially a problem for non-normal modal deontic logics and nonmonotonic deontic logics. Such systems are sometimes weaker than Standard Deontic Logic due to certain restrictions imposed on the logical consequence relation. For instance, different non-normal deontic logics (Van Fraassen, 1973; Chellas, 1980) and nonmonotonic deontic logics (Horty, 1994; Parent and van der Torre, 2018b) have been proposed to deal with moral dilemmas by restricting the aggregation of obligations. In such systems the formula $\mathcal{O}\varphi \wedge \mathcal{O}\neg\varphi \rightarrow \mathcal{O}(\varphi \wedge \neg\varphi)$ is not valid. As pointed out by Van Fraassen (1973) and Horty (1994), letting go of the principle of aggregation is arguably too strong since it also blocks

[16] Recall from Section 1.2.1, that Ross' paradox holds in SDL due to the normality of the deontic operator $\mathcal{O}$ which makes the formula $\mathcal{O}\varphi \rightarrow \mathcal{O}(\varphi \vee \psi)$ an SDL-theorem.

the inference of (A3) from (A1) and (A2). Certain mechanisms may be introduced to the logic to restore some of its inferential power.[17]

Concerning the logic LM, we obtain four possible formalizations of the scenario based on the interpretation of (A1) and (A2) as either an obligation or prohibition. Of the four possible combinations, three are inconsistent:

- $\mathcal{O}(\mathsf{fight} \cup \mathsf{service}/\top)$ and $\mathcal{O}(\overline{\mathsf{fight}}/\top)$ are jointly inconsistent. Namely, both fight and $\overline{\mathsf{fight}}$ are instruments for reducing bad karma, contradicts with the Mīmāṃsā principle that obligations must be non-trivial (cf. axiom A7).

- $\mathcal{O}(\mathsf{fight} \cup \mathsf{service}/\top)$ and $\mathcal{F}(\mathsf{fight}/\top)$ are jointly inconsistent. Namely, by Definition 5.10, the prohibition implies that fight is not simultaneously an instrument for the reduction of bad karma, which contradicts the obligation to perform either fight or $\mathsf{service}$.

- $\mathcal{F}(\overline{\mathsf{fight}}\&\overline{\mathsf{service}}/\top)$ and $\mathcal{O}(\overline{\mathsf{fight}}/\top)$ are jointly inconsistent. The reasoning is similar to the previous item.

- $\mathcal{F}(\overline{\mathsf{fight}} \cap \overline{\mathsf{service}}/\top)$ and $\mathcal{F}(\mathsf{fight}/\top)$ are jointly consistent. The following model, graphically presented in Figure 5.4, demonstrates this.

It remains to check whether the only consistent formalization of the scenario also entails the desired conclusion expressed in (A3), i.e., $\mathcal{F}(\overline{\mathsf{service}}/\top)$. It does.

Proposition 5.1. *The following formula is LM-valid:* $(\mathcal{F}(\overline{\mathsf{fight}}\cap\overline{\mathsf{service}}/\top)\wedge\mathcal{F}(\mathsf{fight}/\top)) \to \mathcal{F}(\overline{\mathsf{service}}/\top)$.

Proof. Take an arbitrary LM-model $\mathfrak{M}$ and a world $w \in W$ of $\mathfrak{M}$ such that (1) $\mathfrak{M}, w \models \mathcal{F}(\overline{\mathsf{fight}} \cap \overline{\mathsf{service}}/\top)$ and (2) $\mathfrak{M}, w \models \mathcal{F}(\mathsf{fight}/\top)$. We reason towards a contradiction by assuming $\mathfrak{M}, w \not\models \mathcal{F}(\overline{\mathsf{service}}/\top)$. This means that either (a) $\mathfrak{M}, w \models \neg\mathcal{I}(\overline{\mathsf{service}}/\mathtt{P}/\top)$ or (b) $\mathfrak{M}, w \models \mathcal{I}(\overline{\mathsf{service}}/\mathtt{R}/\top)$. We consider both cases.

(a) By Definition 5.10 either one of the four clauses (i)-(iv) of instrumentality does not hold.[18] We know that (ii) $\mathfrak{M}, w \models \Diamond_{\mathsf{U}}\top$ and by (1) we know (iii) $\mathfrak{M}, w \models \Diamond_{\mathsf{S}} t(\overline{\mathsf{service}})$. Hence there is a $v \in \mathcal{R}_{\boxed{\mathsf{S}}}(w)$ such that $\mathfrak{M}, v \models t(\overline{\mathsf{service}}) \wedge t(\overline{\mathsf{fight}})$ and by (1) $\mathfrak{M}, v \models \mathtt{P}$. By axiom A6 we know there is a $u \in \mathcal{R}_{\boxed{\mathsf{S}}}(w)$ such that $\mathfrak{M}, u \models \neg\mathtt{P}$. Hence, by (2), we also know $\mathfrak{M}, u \models t(\overline{\mathsf{fight}})$. By contraposition on (1) we obtain $\mathfrak{M}, u \models \neg\mathtt{P} \to \neg(t(\overline{\mathsf{fight}}) \wedge t(\overline{\mathsf{service}}))$ and so by basic LM-reasoning we have $\mathfrak{M}, u \models t(\mathsf{service})$. Consequently, we know clause (iii) $\mathfrak{M}, w \models \Diamond_{\mathsf{S}} t(\mathsf{service})$ also holds. Hence, clause (i) must be violated, i.e., $\mathfrak{M}, w \models \neg\boxed{\mathsf{S}}(t(\overline{\mathsf{service}} \to \mathtt{P})$, which

[17] We refer to the work of Horty (1994) and Parent and van der Torre (2018b) for solutions to this problem. See also Chapter 3 for a discussion of restricted aggregation in deontic STIT logic.

[18] Since the context is $\top$ this condition can be ignored with respect to Definition 5.10.

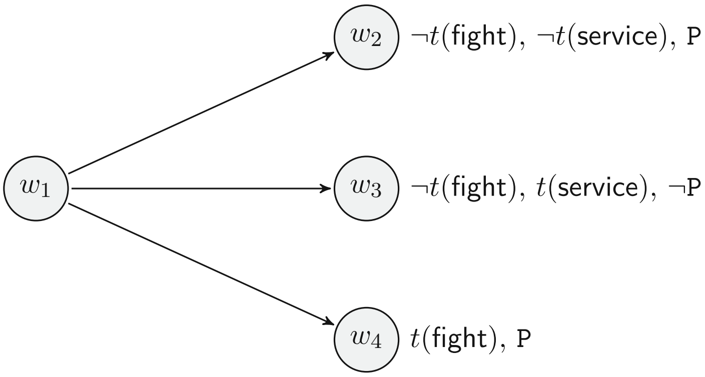

w_2
$\neg t(\mathsf{fight})$, $\neg t(\mathsf{service})$, $\mathtt{P}$
w_1
w_3
$\neg t(\mathsf{fight})$, $t(\mathsf{service})$, $\neg\mathtt{P}$
w_4
$t(\mathsf{fight})$, $\mathtt{P}$

Figure 5.4: An LM-model showing the consistency of the Alternative Service Paradox where both ($\mathsf{A1}$) and ($\mathsf{A2}$) are interpreted as prohibitions. The arrows denote $\mathcal{R}_{\boxed{\scriptstyle S}}$ relations, and the relation $\mathcal{R}_{\boxed{\scriptstyle U}} = W \times W$ is left implicit. The LM-model $\mathfrak{M}$ is defined accordingly: $W = \{w_1, w_2, w_3\}$, $W_{\mathsf{fight}} = \{w_1\}$, $W_{\mathsf{service}} = \{w_1, w_3\}$, $W_{\mathtt{R}} = \emptyset$, $W_{\mathtt{P}} = \{w_1, w_2\}$, $\mathcal{R}_{\boxed{\scriptstyle S}} = \mathcal{R}_{\boxed{\scriptstyle U}} = W \times W$ and V as in Definition 5.5

implies the existence of a $z \in \mathcal{R}_{\boxed{\scriptstyle S}}(w)$ such that $\mathfrak{M}, z \models t(\overline{\mathsf{service}}) \wedge \neg\mathtt{P}$. However, then by (2) we have $\mathfrak{M}, z \models t(\overline{\mathsf{service}}) \wedge t(\overline{\mathsf{fight}}) \wedge \neg\mathtt{P}$ which contradicts with the assumption (1). Hence, all clauses (i)-(iv) are satisfied. Contradiction.

(b) By assumption (1) we know that $\mathfrak{M}, w \models \neg\mathcal{I}(\overline{\mathsf{fight}}\&\overline{\mathsf{service}}/\mathtt{R}/\top)$. However, since clauses (ii)-(iv) of Definition 5.10 are satisfied due to $\mathfrak{M}, w \models \mathcal{I}(\overline{\mathsf{fight}}\&\overline{\mathsf{service}}/\mathtt{P}/\top)$ we know that $\mathfrak{M}, w \models \neg\boxed{\scriptstyle S}((t(\overline{\mathsf{fight}}) \wedge t(\overline{\mathsf{service}})) \to \mathtt{R})$. By the assumption that $\mathfrak{M}, w \models \mathcal{I}(\overline{\mathsf{service}}/\mathtt{R}/\top)$ we know that $\mathfrak{M}, w \models \boxed{\scriptstyle S}(t(\overline{\mathsf{service}}) \to \mathtt{R})$ and thus $\mathfrak{M}, w \models \boxed{\scriptstyle S}((t(\overline{\mathsf{service}}) \wedge t(\overline{\mathsf{fight}})) \to \mathtt{R})$. Contradiction. QED

5.6.4 Jørgensen's Dilemma and Maṇḍana's Reduction

In the 1930s, Jørgenson discussed the following challenges concerning logic and norms, e.g., see (Pigozzi and van der Torre, 2018): logical inference essentially depends on whether sentences (such as premises and conclusions) can be true or false. Norms and imperatives have no truth value. Hence, one cannot justify a norm or imperative through logical reasoning. That is, a logic of norms is impossible. However, valid logical reasoning with norms or imperatives appears possible. This is a dilemma (Hilpinen and McNamara, 2013). The most common approach in deontic logic is to indirectly define reasoning about norms via truth-functional reasoning with propositions about norms. This observation led to the formal distinction between norms and norm propositions (von Wright, 1963a; Makinson, 1999). Then, a deontic formula such as $\mathcal{O}_i\varphi$ is interpreted as a normative proposition, i.e., "it is the case that φ is obligatory for agent i (given the implicit underlying normative code)". The problem of Jørgensen's dilemma for deontic logic is then avoided if we think of deontic logic as a system of logical reasoning with norm propositions. Hilpinen and McNamara (2013) provide an extensive historical discussion of various solutions.

Interestingly, Maṇḍana's deontic theory offers a novel contribution to dealing with Jørgenson's dilemma. Commands are for Maṇḍana descriptive sentences about the world, namely, about relations between actions and particular states of affairs that result from those actions. Consequently, Maṇḍana's theory reduces norms (or normative statements, for that matter) to truth-functional propositions about the world. In other words, Maṇḍana's deontic theory provides an alternative solution to the dilemma.

In sum, the discussed deontic puzzles do not lead to inconsistencies in LM. These encouraging results may be due to the depth of Maṇḍana's deontic theory. Although some of the various formal interpretations of contrary-to-duty scenarios were inconsistent, most interpretations yielded consistent formalizations. Maṇḍana's deontic theory addresses the paradoxes in accordance with a common solution strategy adopted in modern deontic logics. Namely, the adaptation of a logic of actions with an Andersonian reduction (Bartha, 1993; Castañeda, 1981; Giordani and Canavotto, 2016; Meyer, 1988; Meyer et al., 1994). This is a surprising convergence for a philosophical approach whose foundations lie millennia back. We refer to (van Berkel et al., 2022a) for further analysis of the paradoxes.

5.7 Related Work and Future Research

Action Logics. Since Maṇḍana's elementary concepts are actions and outcomes, we adopted a PDL-like language (Fischer and Ladner, 1979; Meyer, 1988). For our purposes, a minimal action language sufficed using negation, disjunction, and conjunction. Despite its simplicity, this language allows for notions of instruments that, for instance, take actions as preconditions. Since LM is similar to the Logic and Agency and Norms in Chapter 4, we refer to that chapter for a more extensive discussion of action logics.

Deontic Reductions: Sanction, Violation, and Ideality. The reduction of Vedic commands offered by Maṇḍana has some striking similarities with what is known as the Andersonian-Kangerian reduction of (Standard) Deontic Logic. In the early days of deontic logic, Anderson and Moore (1957) argued for a reduction of deontic modalities to the (then) better understood alethic modalities, motivated by the striking similarity between the two modalities, i.e., necessity and obligation behave as universal operators, whereas possibility and permissibility behave as existential operators.[19] Anderson proposed a treatment of obligations as statements concerning necessity and violation: "φ is obligatory if $\neg\varphi$ necessary imply a sanction".[20] Later, Castañeda (1972) argued that a conceptual modification of 'sanctions' to 'violations' solves several philosophical issues with the

[19] This correspondence was already noticed and discussed by von Wright (1951).

[20] As pointed out by Ciabattoni et al. (2021), Anderson's reduction is similar to the stance taken by Kelsen in his theory of law. There, Kelsen argues that "the legal order [. . .] prohibits a certain behavior by attaching to it a sanction or [. . .] it commands a behavior by attaching a sanction to the opposite behavior" (as quoted by Ciabattoni et al. (2021) on page 141).

reduction. It is well-known that Standard Deontic Logic—a non-normal modal logic interpreting obligation as a KD modality—can be translated into Anderson's logic, e.g., see (Parent and van der Torre, 2018a). However, we recall that Anderson's logic (1957) is more expressive than Standard Deontic Logic, e.g., it can characterize the principle of deontic contingency (cf. page 189). The reduction was later adopted by Meyer (1988) to an action setting: "an action Δ is obligatory if all performances of its complement $\overline{\Delta}$ lead to a violation". D'Altan et al. (1996) use both approaches to provide a uniform setting for discussing 'ought-to-be' versus 'ought-to-do'. In Chapter 4 (van Berkel et al., 2020), the reduction-setting was to the inclusion of norms of instruments.

Similarly, Maṇḍana can be regarded as a reductionist of deontic reasoning: he regards every Vedic command as a statement concerning instruments, that is, actions leading to different results. The results are of three types: states of affairs, sanctions, and rewards. This trichotomy proposed by Maṇḍana has also, independently, been proposed in the history of deontic logic: Kanger (1971) proposed the inclusion of a positive constant "what morality prescribes" as a means to identifying obligations, whereas prohibitions would be defined in terms of sanctions. There, an obligation is not identified in terms of sanctions but in terms of 'what morality prescribes': "φ is obligatory if it is necessary that morality prescribes φ". Formally, the obligation is defined as $\mathcal{O}\varphi := \Box(\texttt{w\_m\_p} \rightarrow \varphi)$ (where $\texttt{w\_m\_p}$ is a constant denoting 'what morality prescribes'). A significant difference between Kanger's approach and Maṇḍana's, is that the former takes φ as a *necessary* condition for the 'ideal world' whereas for Maṇḍana φ is a *sufficient* condition for 'reducing bad karma'. See the work of Glavaničová and Pascucci (2021) for another reductionist approach to deontic modalities. There, (conditional) obligations are defined using an ideality constant and an ideality and subideality modality. See the work of Lomuscio and Sergot (2003) for the use of "green" constants identifying permissible states.

Formalizations of the Śyena. In this chapter, we discussed a formal analysis of Maṇḍana's solution to the Śyena controversy. The analysis was initially presented in (van Berkel et al., 2021a), which also contains formal analyses of the solutions provided by Prabhākara and Kumārila. Guhe (2021), provides a formal analysis of the solution of the Navya-Nyāya school. We briefly discuss these analyses here.

Prabhākara's solution. Prabhākara does not deontically distinguish between obligations and elective commands. For him, all prescriptions are obligations proper. Prabhākara's solution to the Śyena is a case of CTD reasoning. Namely, the command to perform the Śyena is an obligation activated when a specific violation occurs. The violation is the desire to cause an enemy's death. It must be noted that Prabhākara takes desires as irreversible decisions. Thus, for Prabhākara, the desire to kill amounts to a decision to kill. In ideal circumstances, no living being would desire the death of their enemy and, therefore, would not be eligible to perform the Śyena. By contrast, once the subideal eligibility condition of desiring the death of their enemy is satisfied, the performance of the Śyena becomes obligatory. Prabhākara's solution to the Śyena was consistently formalized by Ciabattoni et al. (2015) and a more refined analysis was provided in (van Berkel et al., 2021a). The proposed logic for Prabhākara's deontic theory consists of a propositional logic extended with a global modality $\boxed{\mathsf{U}}$ and two dyadic deontic operators $\mathcal{O}(./.)$ and $\mathcal{F}(./.)$ for obligation and prohibition, respectively. The modal logic is a non-normal modal (Chellas, 1980).

Kumārila's solution. Although the deontic theories of Prabhākara and Kumārila are quite similar, the two authors have different solutions to the Śyena controversy. Kumārila's solution relies on the distinction between obligations and elective commands. In his deontic theory, elective commands have no deontic force. Elective commands are mere Vedic recipes that provide suitable means to achieve specific goals. Consequently, in case of a conflict with a prohibition, an agent just can refrain from performing the elective sacrifice, thus avoiding a violation of the prohibition. Likewise, in the case of Śyena, one simply does not perform it. Kumārila's solution to the Śyena was consistently formalized and analysed in (van Berkel et al., 2021a). The proposed logic extends the logic for Prabhākara's deontic theory (see above) with an additional dyadic modal operator $\mathcal{E}(./.)$ representing elective commands. This additional modality is characterized by weaker properties than those for the other deontic operators. The only requirement imposed on elective commands is self-consistency.

The Navya-Nyāya solution. Guhe (2021) formally investigates Gaṅgeśa's solution to the Śyena controversy, Gaṅgeśa belongs to the Navya-Nyāya school. Guhe discusses the school's account of obligations, permissions, and prohibitions. In contrast to common Mīmāṃsā, injunctions have an inherently teleological meaning for the Navya-Nyāya. Namely, in order to induce an agent to obey (Vedic) commands, the presence of the (un)desired expected effect is postulated. The Navya-Nyāya also maintain a reduction of commands. The approach adopted by the Navya-Nyāya is strikingly akin to Maṇḍana's approach. One of the main differences is that, for Maṇḍana, obligations can be neglected without additional sanctions, whereas for the Navya-Nyāya, the neglect of an obligation causes the accumulation of bad karma. Gaṅgeśa's solution to the Śyena is that even though it is prescribed, performing the Śyena is inappropriate for virtuous agents. A virtuous agent pursues any action that ends "in a deontically perfect world" (Guhe, 2021, p.28). The performance of Śyena is not part of any deontically perfect world.

For its formalization, Guhe (2021) adopts the dynamic deontic logic ADL, as developed by Giordani and Canavotto (2016). It is a deontic action logic in the Andersonian tradition (see Chapter 4 for a discussion of this logic). Guhe (2021) adopts constants referring to deontically perfect worlds and to sanctions. An instrument is a formula of the form $[\Delta]\varphi$, where $[\Delta]$ is a primitive modality (cf. $\boxed{\mathsf{s}}(t(\Delta) \to \varphi)$ in the language of LM). It is clear from Definition 5.10 that Maṇḍana's conception of instrumentality is more involved than the one adopted by Guhe (2021) (cf. the contingency and meaningfulness requirements in Section 5.4). The Navya-Nyāya solution to the Śyena controversy can be consistently formalized in ADL due to the use of two independent deontic operators. One of the important differences between our formalization and the one by Guhe (2021) is that our logic is solely constructed from deontic properties found in Mīmāṃsā source texts. The logic used by Guhe (2021) was developed independently of the former's aim, namely, by

Giordani and Canavotto (2016). This may lead to questions concerning the validity of the formal analysis since certain logical consequences of the formalization may be due to the underlying logic ADL.

The debate on Śyena is not limited to the above authors. The topic has been thoroughly investigated in Sanskrit philosophy for more than two millennia. For an in-depth analysis of various solutions to the controversy, we refer to (van Berkel et al., 2022a).

Related Mīmāṃsā work. Ciabattoni et al. (2015) were the first to formalize Mīmāṃsā reasoning in logic. Their system is called "basic Mīmāṃsā Deontic Logic" (bMDL). Although entirely based on Mīmāṃsā principles, the necessity-free fragment of this logic is, in fact, identical to the dyadic version of the non-normal deontic logic MD (Chellas, 1980) (also see the work by Freschi et al. (2019) and Lellmann et al. (2021)). The logic bMDL captures the concept of obligation in common Mīmāṃsā—encompassing both fixed and occasional duties—but adopts obligation as its only deontic operator. The logics formalizing the deontic theories of Prabhākara and Kumārila, called LPr^+ and LKu^+ respectively, were introduced in (van Berkel et al., 2021a). Preliminary versions of these logics were presented by Lellmann et al. (2021). There, a nonmonotonic sequent-style proof system was developed for Prabhākara's deontic theory. The system captures a Mīmāṃsā-based interpretation of a contemporary defeasible principle known in AI as *specificity*, e.g., see (Horty, 2012). The two logics LPr^+ and LKu^+ extend (a variant of) bMDL with prohibitions and, in the case of Kumārila, also with elective duties. In (van Berkel et al., 2022a), these two logics were refined and extended with a restriction aggregation principle (i.e., **P4** discussed in Section 5.4).

Talmudic logic and more. Mīmāṃsā inspired logics are not the only attempts at modeling ancient theories of normative reasoning. It goes beyond the scope of this chapter to discuss works on non-deontic formalizations of ancient theories. One work that deserves mention due to its closeness to the approach followed in this chapter is Talmudic deontic logic, developed by Abraham et al. (2011). The first thing to note is that obligation and prohibition are not interdefinable in the Talmudic interpretation. Furthermore, the logic underlying their modal logic is intuitionistic. Consequently, $\neg\neg\mathcal{F}\varphi$ is not equivalent to $\mathcal{F}\varphi$ since "[t]he first is only a weak prohibition, a recommendation for good behavior in the eyes of God, while the second is a full-fledged strong prohibition" (Abraham et al., 2011, p.120). Another distinctive characteristic of their formalism is that deontic conflicts are not resolved through applying general principles of reasoning, such as for the Mīmāṃsā, but through rabbis making decisions based on principles. Like Maṇḍana, in Talmudic deontic logic, the fulfillment of an obligation leads to a reward with no punishment for failing to fulfill an obligation, and the fulfillment of a prohibition yields no reward, whereas violating it causes a sanction.

* * *

In conclusion, we analyzed the deontic theory of Maṇḍana, one of the central authors of the Mīmāṃsā school. His theory reduces all Vedic commands to statements about actions as instruments leading to specific results. We provided a sound and complete logic LM capturing this deontic reduction (Objective 1). We employed the logic to improve our understanding of Maṇḍana's theory and to evaluate whether Maṇḍana's reduction preserves the validity of central reasoning principles used by the Mīmāṃsā in general (Objective 2). Thereafter, we provided a logical analysis of Maṇḍana's solution to the Śyena controversy (Objective 3) and evaluated the logic LM on some well-known benchmark puzzles from the deontic logic literature (Objective 4). The present results only scratch the surface of the research opportunities offered by formal approaches to the study of Mīmāṃsā deontic reasoning.

Part III

Argumentation and Normative Reasoning

CHAPTER 6

Deontic Explanations

In the previous chapters, we used logical methods to reason about agents in the context of norms. This chapter investigates whether we can make formal normative reasoning more accessible *to* agents: we focus on *explanations*. The fundamental role of norms is to motivate, guide, and limit the choices made by agents (Chapter 1). It often does not suffice for a real-life agent to know that an obligation applies to her. To motivate compliance, the agent must understand why she is required to behave in a particular way (especially if she disagrees with her alleged duty). We call answers to such why questions *deontic explanations*. These explanations not only lead to a better understanding of normative reasoning, they also motivate appropriate conduct, enhancing compliance and improving collaboration (Chopra et al., 2018). Moreover, the importance of such explanations increases with the continuing development of intelligent autonomous systems that must comply with normative codes; cf. (Liao et al., 2019). For instance, in applying formal normative reasoning to autonomous cyber-physical systems such as self-driving cars (Shea-Blymyer and Abbas, 2021), we can better understand the behavior of those systems when we know why particular normative choices are made. Currently, there is no formal system that facilitates deontic explanations.

The present chapter, based on the results in (van Berkel and Straßer, 2022), introduces Deontic Argumentation Calculi (DAC for short), which are sequent-style proof systems tailored to the construction of deontic explanations. With this, we aim to lay a foundation for the study of explanations in deontic logic.

Deontic Logic and Reasons. When answering the question as to *why* an obligation holds, one must state *reasons*. However, it often does not even suffice to know why a specific obligation holds without knowing why other obligations *to the contrary* do not hold. Especially in the light of (potential) conflicts, answers to such contrastive why questions become crucial. For example, to understand why "I am permitted to overtake on the left, *despite* having to drive on the right" I must know how the first norm relates to the second. In this case, the first norm is an exception that renders the latter *inapplicable* in the context of “overtaking another vehicle”. Unfortunately, common approaches in deontic logic do not provide means for making explicit the reasons why certain obligations are not derivable. Despite their central role in ethics and explanation (Brunero, 2018), a general lack of explicit modeling of reasons in formal systems was recently identified by Nair and Horty (2018) (a notable exception being the work of Horty (2012)).

Objective 1. *Develop a deontic language in which reasons are explicitly formalized.*

Complex normative systems often require reasoning with normative conflicts, exceptions, preferences, and a variety of resolution mechanisms (Delgrande, 2020; Gabbay et al., 2013; Tosatto et al., 2012). These challenges can be effectively addressed using nonmonotonic reasoning (Nute, 1997). (We refer to Hilpinen and McNamara (2013) for a discussion of alternative approaches.) In this respect, the Input/Output formalism (I/O, for short) as developed by Makinson and van der Torre (2000) is particularly promising (see Chapter 1). The I/O formalism facilitates reasoning with norms and contains various mechanisms to defeasibly detach obligations from norms in a given context. In particular, *constrained I/O logics* (Makinson and van der Torre, 2001) have been employed to nonmonotonically reason with deontic conflicts, contrary-to-duty scenarios, and exceptions. One advantage of the I/O approach is that it formalizes norms as reasons. To illustrate, let (p, q) be a norm expressing “given fact p, it is obligatory that q”. Then, in the context of fact p, the norm (p, q) is a reason why it is obligatory that q. Nevertheless, I/O leaves some of the challenges mentioned above unaddressed. For instance, nonmonotonicity is captured in the traditional constrained I/O approach by considering maximally consistent sets of norms. In this approach, norms are not part of the object language, and reasoning *about* norms takes place on a meta-level. Thus, one cannot readily use the framework to explicitly reason with reasons, generating explanations accordingly.

Objective 2. *Develop formal calculi for Input/Output logics, fully integrating meta-reasoning about norms into the object language of the calculi. In particular, the calculi must generate transparent arguments that provide reasons why obligations hold and why certain norms are inapplicable.*

We address the above two objectives by introducing a class of rule-based sequent-style proof systems called *Deontic Argument Calculi* (DAC for short). The calculi internalize the meta-reasoning of I/O logics and generate arguments that provide direct and explicit reasons for obligations as well as for the inapplicability of norms. We adopt a sequent-style approach due to its high modularity, suitability for proof-theoretic analysis, and intuitive construction of derivations as trees (Negri et al., 2008) (also see page 10). Our approach belongs to sequent-based tradition to logical argumentation (Arieli and Straßer, 2015; Arieli et al., 2022b; Straßer and Arieli, 2014; Straßer and Arieli, 2015).

Formal Argumentation. Over the past decades, formal argumentation (Dung, 1995) has proven to be a unifying framework for the representation of large classes of nonmonotonic logics (Arieli et al., 2021). The central concept of this field is that of an argumentation framework. It consists of a set of arguments together with an attack relation defining conflicts between these arguments. The idea of defeasibility is then captured in terms of counterarguments attacking an initial argument. Formal argumentation provides both a natural and a transparent model of conflicts and their resolution. Consequently, it serves as a promising basis for tackling the central challenges of normative reasoning discussed above. In recent years, argumentative representations of deontic logics have attracted increasing interest (Beirlaen et al., 2018; Governatori et al., 2018; Liao et al., 2018; Peirera et al., 2017; Pigozzi and van der Torre, 2018; Straßer and Arieli, 2015). In this chapter, we set out to combine the advantages of the I/O formalism with those of formal argumentation. Namely, on the one hand, I/O is a highly expressive and robust framework with more than two decades of development, e.g., see (Parent and van der Torre, 2013; Parent and van der Torre, 2018b). On the other hand, I/O does not provide the level of transparency that comes with an explicit representation of conflicts in formal argumentation.

Objective 3. *Provide a formal characterization of the nonmonotonic inference relation of standard Input/Output approaches in formal argumentation frameworks instantiated with* DAC *arguments.*

By addressing this objective, we open the door to applying existing explanation methods developed for formal argumentation to I/O reasoning.

Formal Explanations. Explanation is gradually taking up a more central position in AI (Miller, 2019). Argumentative approaches are promising in this respect due to their intuitive representation of and closeness to human reasoning practices (Mercier and Sperber, 2011). In fact, the study of explanation is also gaining traction in formal argumentation (Arieli et al., 2022b; Borg and Bex, 2021; Čyras et al., 2021; Fan and Toni, 2015a; Liao and van der Torre, 2020; Vassiliades et al., 2021). Explanations can be seen as specific types of arguments (Šešelja and Straßer, 2013). An argument shows that some statement is correct, whereas an explanation additionally shows *why* and *how* this statement is correct. The former's aim is justification, whereas the latter's aim is understanding. To illustrate, a derivation provided by a proof system is a certificate that justifies that something is derivable. However, it may not suffice as an explanation: a derivation is not necessarily transparent, may contain redundant and irrelevant steps, and may not be understandable by an agent not versed in proof theory. Although certificates of derivability (such as derivations generated by a proof system) are essential for justification, explanation requires more than justification.

There are two essential tiers of explanation related to formal argumentation (Johnson, 2000). First, there is explanation on the level of an argument (Toulmin, 1958), i.e., an argument is not just a logical derivation; it has a structure with additional qualifying information, e.g., warrants, strengths, defaults, and supports. We call this *internal explanation* (cf. the illative tier of Johnson (2000)). Second, there is explanation on the level of argument interaction. Here, explanation of an argument is given by other arguments through notions of attack and defense. We call this *external explanation* (cf. the dialectic tier of Johnson (2000)). Ideally, good explanations are a mix of both, e.g., contrastive explanations (Stepin et al., 2021). However, so far, little to no formal work has been done in the intersection of explanation and normative reasoning. This brings us to our final objective.

Objective 4. *Employ the developed calculi to generate deontic explanations using tools from formal argumentation.*

We illustrate the utility of our approach using the notion of *related admissibility* developed by Fan and Toni (2015a). In particular, we use this notion to explain why some obligations hold *despite* certain norms to the contrary.

Contributions. In this article, we take the formal argumentation path to deontic explanation and introduce a class of *Deontic Argumentation Calculi* (DAC). Our focus is on both internal and external explanations. Deontic Argumentation Calculi accommodate explanation in a variety of ways:

1. We use labels on formulae to make the presentation transparent on the object level, i.e., we can syntactically distinguish between facts, obligations, and constraints without "burdening" the logics with modalities; cf. (Makinson and van der Torre, 2000; Parent and van der Torre, 2018b).
2. We internalize some of the meta-reasoning in the I/O formalism by referring to the inapplicability of norms on the object language level.
3. We represent explicit *reasons* in the premises of arguments such that, by presenting arguments for a conclusion, the reasons are immediate.
4. We provide an additional admissible DAC rule that makes DAC arguments *relevance-aware*, i.e., premises only bear reasons directly relevant for the derivation of the conclusion.

All of the above points increase the *self-explanatory character* of arguments, for instance, in contrast to classical argumentation; see (Arieli et al., 2021). In sum, our calculi generate both arguments that provide explicit reasons for obligations and arguments that defeat other arguments by giving explicit reasons for why certain norms are inapplicable. The second type of argument concerns the defeasibility of normative reasoning. In other words, our approach accommodates internal and external explanations. Consequently, it can be used to answer non-contrastive "why A?" and contrastive "why A rather than B?" questions. The possibility to reason about the inapplicability of norms on the object language level distinguishes our work from other proof-theoretic approaches to Input/Output logic, e.g., by Lellmann (2021) and Straßer et al. (2016).

The technical contribution of this chapter consists of two types of completeness results:

1. We show adequacy between DAC and a significant class of monotonic I/O consequence relations.

2. We prove that formal argumentation frameworks instantiated with DAC arguments characterize a large class of nonmonotonic I/O logics.

These contributions make our work the first to characterize a significant class of (non)monotonic I/O logics, including all original logics by Makinson and van der Torre (2000; 2001). The addition and removal of rules in DAC correspond to different I/O logics, and our calculi enjoy a modularity particularly suitable for expansions. Moreover, DAC is modular for a large class of underlying base logics (i.e., not just classical and intuitionistic logic). Last, our work enhances the scope of previous representation results in formal argumentation concerning systems based on maximally consistent sets (Arieli et al., 2021).

The formalism developed in this chapter provides the foundation for a more extensive investigation of explanation in the context of normative reasoning. In Section 6.8, we discuss promising research directions in this respect.

Differences. Deontic Argumentation Calculi were introduced in (van Berkel and Straßer, 2022). The additional contribution of this chapter is the inclusion of relevance rules. These rules can be employed in two ways: They can be used to generate smaller DAC-instantiated argumentation frameworks, excluding arguments with irrelevant reasons, or they can be exploited to generate arguments that attack irrelevant arguments.

Outline of this work. We provide basic terminology and two running examples in Section 6.1 and recap I/O logic in Section 6.2. In Section 6.3, we present our Deontic Argumentation Calculi. We show soundness and completeness between traditional I/O proof systems and Deontic Argumentation Calculi in Section 6.4. In Section 6.5, we define formal argumentation frameworks instantiated with DAC-arguments and illustrate how existing approaches to explanation in formal argumentation can be used in our setting. Soundness and completeness between DAC-instantiated argumentation frameworks and nonmonotonic inference in I/O logics are shown in Section 6.6. In Section 6.7, we extend DAC with relevance rules. In Section 6.8, we discuss related and future work.

6.1 Basic Terminology and Benchmark Examples

Developments in deontic logic are driven by challenging examples and paradoxes (Chapter 1). In this section, we introduce basic terminology by considering two examples. Here, we focus on *contrary-to-duty* reasoning and *deontic dilemmas*. Both can be effectively addressed using nonmonotonic reasoning (Nute, 1997; Parent and van der Torre, 2018b).[1]

[1] For alternative approaches, see the overview by Hilpinen and McNamara (2013).

The language $\mathcal{L}$ employed throughout this chapter is given in Definition 6.1. To increase the transparency of our formal representation, we *label* formulae of $\mathcal{L}$.

Definition 6.1 (The Language $\mathcal{L}$)**.** *Let* $\mathsf{Atoms} = \{p, q, r, \dots\}$ *be a denumerable set of propositional atoms. The language* $\mathcal{L}$ *is defined via the following BNF grammar:*

$$\varphi ::= p \mid \top \mid \bot \mid \neg\varphi \mid \varphi \vee \varphi \mid \varphi \wedge \varphi \mid \varphi \rightarrow \varphi$$

with $p \in \mathsf{Atoms}$.

Let the labelled *language* $\mathcal{L}^i$ *be defined as* $\mathcal{L}^i := \{\varphi^i \mid \varphi \in \mathcal{L}\}$ *for* $i \in \{f, o, c\}$*. We say that* $\mathcal{L}^f$ *is the language expressing* facts*,* $\mathcal{L}^o$ *is the language expressing* obligations*, and* $\mathcal{L}^c$ *is the language expressing* constraints*. Last, let* $\mathcal{L}^n := \{(\varphi, \psi) \mid \varphi, \psi \in \mathcal{L}\}$ *be the language of* norms *(where the superscript* n *refers to norms).*

All connectives are primitive in order to be modular with respect to a large class of propositional base logics. We come back to this in Section 6.2. We use $p, q, r, \dots$ (possibly indexed) for atoms, and reserve $\varphi, \psi, \theta, \dots$ (possibly indexed) for arbitrary formulae of $\mathcal{L}$. A formula $\varphi^f \in \mathcal{L}^f$ expresses “it is a fact that φ”, formula $\varphi^o \in \mathcal{L}^o$ states that “it is obligatory that φ”, and a formula $\varphi^c \in \mathcal{L}^c$ denotes that “φ is a constraint”.[2] Moreover, following Makinson and van der Torre (2000), we take a pair of propositional formulae (φ, ψ) to represent a *norm*, i.e., “given fact φ, it is obligatory that ψ”.

In contrast to the previous chapters, in this chapter, we adopt labels instead of modalities to denote the role of the various formulae involved in normative reasoning. The main reason for doing so is our aim to provide a novel proof-theoretic characterization of a class of I/O logics while remaining as faithful to the original approach as possible. Labels provide a transparent way to representing the various roles propositional formulae play in normative reasoning.[3]

We work with knowledge bases of the type $\langle \mathcal{F}, \mathcal{N}, \mathcal{C} \rangle$, where $\mathcal{F} \subseteq \mathcal{L}^f$ constitutes the *factual context*, $\mathcal{N} \subseteq \mathcal{L}^n$ denotes a *normative code*, and $\mathcal{C} \subseteq \mathcal{L}^c$ represents the *constraints* with which output must be consistent. The basic idea is that facts (input) trigger norms from which obligations are detached (output). Moreover, constraints control the output to ensure consistency. The above is in the spirit of constrained I/O logic by Makinson and van der Torre (2001).

An *explanatory argument* is an argument stating reasons for a conclusion. Suppose we have a single fact $\mathcal{F} = \{p^f\}$, a norm system $\mathcal{N} = \{(p, q), (r, s)\}$, and no constraints, then an argument concluding that q is obligatory is of the following form,

$$\underbrace{\overbrace{p^f, (p, q)}^{\text{reasons...}} \quad \Rightarrow \quad \overbrace{q^o}^{\text{for}}}_{\text{argument}}$$

[2] Since we do not allow for formulae with mixed labels, we can safely omit brackets concerning the use of labels, e.g., we write $\neg\varphi^f$ instead of $(\neg\varphi)^f$.

[3] Modal representations of some I/O logics are available, e.g., see (Makinson and van der Torre, 2000; Lellmann, 2021).

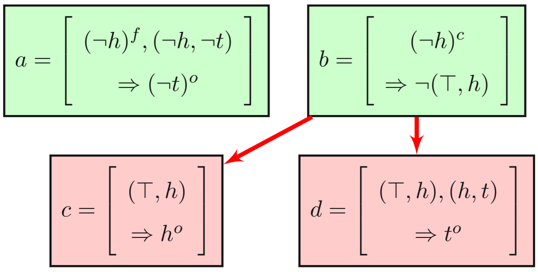

$a = \left[\begin{array}{c}(\neg h)^f, (\neg h, \neg t)\\ \Rightarrow (\neg t)^o\end{array}\right]$
$b = \left[\begin{array}{c}(\neg h)^c\\ \Rightarrow \neg(\top, h)\end{array}\right]$
$c = \left[\begin{array}{c}(\top, h)\\ \Rightarrow h^o\end{array}\right]$
$d = \left[\begin{array}{c}(\top, h), (h, t)\\ \Rightarrow t^o\end{array}\right]$

Figure 6.1: Defeasible normative reasoning examples: The Chisholm scenario (Example 6.1). Arrows denote defeat relations between arguments, relative to the constraint set $\mathcal{C}' = \{\neg h^c\}$ (Example 6.2).

The left-hand side (lhs) gives reasons for the conclusion on the right-hand side (rhs). Alternatively, one can think of the lhs as the *explanans* and the rhs as the *explanandum* (Šešelja and Straßer, 2013). The arrow ⇒ denotes that the rhs "follows from" the lhs. We make this formally precise in Section 6.3. Furthermore, in explanatory reasoning, we do not want arguments to contain information irrelevant to explaining the output. For instance, in the argument $p^f, (p, q), (r, s) \Rightarrow q^o$ the norm (r, s) is irrelevant for the conclusion q^o. As a desideratum, explanatory arguments must be relevance-aware. We address the desideratum at the end of this chapter in Section 6.8.

Example 6.1 (Chisholm scenario (1963), Figure 6.1)**.** *Billy is obligated to go and* help *her neighbors $(\top, h)$. A norm $(\top, h)$ with as a precondition the tautology $\top$ is triggered by default, that is, even by an empty factual context. Furthermore, Billy knows that if she goes to help, she must* tell *them she goes (h, t). Now, if Billy does not go, she ought not to tell them she goes $(\neg h, \neg t)$. It turns out that Billy does not go to help $\neg h^f$. Clearly, Billy has violated her obligation to go and help. Let the knowledge base be $\mathcal{F} = \{\neg h^f\}$ and $\mathcal{N} = \{(\top, h), (h, t), (\neg h, \neg t)\}$. Figure 6.1 presents arguments a, c, and d that can be constructed*[4] *from the knowledge base (we explain the meaning of b and its corresponding arrows in Example 6.2); e.g., in argument a the reasons for not telling $\neg t^o$ are the fact $\neg h^f$ and the norm $(\neg h, \neg t)$. What must Billy do in this* contrary-to-duty *scenario? The desired answer is that she ought not to tell the neighbors she goes $\neg t^o$. Formalizations of this scenario cause problems for Standard Deontic Logic (see page 13), e.g., both telling t and not telling $\neg t$ become obligatory.*

Arguments do not only provide reasons in support of a concluded obligation (i.e., the *explanandum*) but also defend them from potential *defeaters*. Different types of defeat are possible: a rebutting defeat attacks conclusions, an undermining defeat attacks premises, and an undercutting defeat attacks the application of rules (Modgil and Prakken, 2014). Not all types of defeat are suitable for the kind of explanation we have in mind. For

[4] For now, it suffices to leave the construction of such arguments implicit. In Section 6.3, where we introduce the envisioned Deontic Argumentation Calculi, we demonstrate how these arguments can be derived using the calculi.

instance, a rebutting defeat opposes the conclusion of an argument without pinpointing the reason as to why. In contrast, attacks on reasons—i.e., *undercutting* defeats—are arguments that express *which reasons* are *inapplicable* in the light of a given context, i.e., other reasons, facts, and constraints. We adopt undercuts since they are more transparent about attacks. Recall that constraints are consistency requirements and suppose $\mathcal{C} = \{\neg q^c\}$. Then, a defeating argument

$$p^f, \neg q^c \Rightarrow \neg(p, q)$$

expresses that if the output is to be consistent with the constraint $\neg q^c$, in context p^f, the norm (p, q) cannot be consistently asserted as a reason (since it would detach q^o). Hence, $\neg(p, q)$ expresses that this norm is *inapplicable* given $\mathcal{F}$ and $\mathcal{C}$. An argument concluding $\neg(p, q)$ defeats all arguments that appeal to (p, q) as a reason. An argumentation framework is then a set of arguments with defeat relations holding between them.

Example 6.2 (Example 6.1 cont.)**.** *We want to know what Billy must do in the light of her violation $\neg h^f$. Thus, we impose the constraint that the output must be consistent with the fact that Billy does not help $\mathcal{C} = \{\neg h^c\}$ (i.e., $\mathcal{C} = \mathcal{F}$ modulo relabelling). This constraint gives us the argument $b : \neg h^c \Rightarrow \neg(\top, h)$ expressing that given consistency requirement $\neg h$, the norm $(\top, h)$ may not be asserted as a reason (it would output the inconsistent h^o).*[5] *This argument serves as a defeater of any argument that appeals to $(\top, h)$ in its reasons. In this case, this includes arguments c and d. See the defeat arrows in Figure 6.1. So, that Billy ought not to tell $\neg t^o$ is explained by argument a together with the fact that arguments c and d concluding helping h^o, respectively telling t^o, cannot be defended in view of b. Namely, c and d employ reasons that are inapplicable given $\mathcal{C}$.*

What makes this approach more transparent than traditional formalisms such as Input/Output logic (Makinson and van der Torre, 2000) and logical argumentation (Arieli and Straßer, 2019) is the use of labels to indicate different types of information (factual, obligations, constraints), the internalized meta-reasoning about the inapplicability of norms, the argumentative representation of reasons (lhs of ⇒) for the conclusion (rhs of ⇒), and the argumentation framework revealing the contrastive explanatory dimension of defeasible reasoning. In Figure 6.1, the question "why shouldn't Billy help, *despite* argument c?" is answered by "since argument b attacks c and b is not attacked".

Let us now turn to another type of deontic scenario that proves challenging for formal systems of normative reasoning: the *deontic dilemma* (cf. Chapter 3, page 102).

Example 6.3 (Deontic Dilemma, Figure 6.2)**.** *Maxwell and Joan are two colleagues. Joan built a shed and borrowed Max's hammer for the job. She is under the obligation to* return *the hammer to Max $(\top, r)$. At some moment, a deranged Max is standing at the door asking Joan to give her the hammer, he claims, "in order to bang someone's head*

[5] Although in monadic deontic logics "not helping" and "it is obligatory to help" are consistent, in constraint I/O reasoning the former proposition is used to *block* undesirable consequences like the conclusion "it is obligatory to tell".

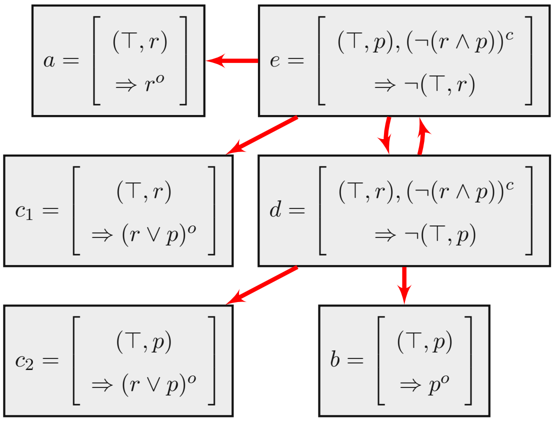

$a = \left[\begin{array}{c}(\top, r)\\ \Rightarrow r^o\end{array}\right]$
$e = \left[\begin{array}{c}(\top, p), (\neg(r \wedge p))^c\\ \Rightarrow \neg(\top, r)\end{array}\right]$
$c_1 = \left[\begin{array}{c}(\top, r)\\ \Rightarrow (r \vee p)^o\end{array}\right]$
$d = \left[\begin{array}{c}(\top, r), (\neg(r \wedge p))^c\\ \Rightarrow \neg(\top, p)\end{array}\right]$
$c_2 = \left[\begin{array}{c}(\top, p)\\ \Rightarrow (r \vee p)^o\end{array}\right]$
$b = \left[\begin{array}{c}(\top, p)\\ \Rightarrow p^o\end{array}\right]$

Figure 6.2: Defeasible normative reasoning examples: A deontic dilemma (Example 6.3). Argument e defends $\{b, c_2, e\}$, whereas argument d defends $\{a, c_1, d\}$.

in" (Maxwell has an unfortunate temper). Joan feels she has the duty to prevent *Max from hurting anyone* $(\top, p)$*. Furthermore, the constraint is that Joan cannot both return the hammer and prevent harm from being done with it* $\neg(r \wedge p)^c$*. What should Joan do? This scenario illustrates a* deontic dilemma *and can be ultimately traced back to Plato (Lemmon, 1962). The knowledge base is defined as* $\mathcal{F} = \emptyset$*,* $\mathcal{N} = \{(\top, r), (\top, p)\}$*, and* $\mathcal{C} = \{\neg(r \wedge p)^c\}$*. The arguments that can be constructed are presented in Figure 6.2. The two defeating arguments, d and e, express that given the constraints, one of either two norms cannot be asserted. Furthermore, in this example, we suppose we reason classically, which means that* p *and* r *both entail* $p \vee r$*; cf. arguments* c_1 *and* c_2*.*

Intuitively, in Figure 6.2, the defensible set $\{a, c_1, d\}$ justifies the obligation that Joan ought to return the hammer, whereas $\{b, c_2, e\}$ does this for the prevention of harm being done. Likewise, one can give an explanation for the floating conclusion $(r \vee p)^o$ in Figure 6.2, by arguing that in every defensible stance *either* c_1 *or* c_2 is selected; cf. disjunctive response in Chapter 3. A floating conclusion is a formula that is derivable through several conflicting arguments (Straßer and Antonelli, 2019), e.g., the arguments c_1 and c_2 both conclude $(r \vee p)^o$ but belong to the two conflicting sets of arguments $\{a, c_1, d\}$, respectively $\{b, c_2, e\}$. However, following a more skeptical reasoning style, one can argue why $r \vee p$ is not obligatory since there is no single argument concluding $(r \vee p)^o$ that is selected in *every* defensible set. Defeasible reasoning by means of argumentation gives rise to various reasoning styles, including those mentioned above. The framework presented in Section 6.5 can disambiguate between them.

All notions discussed in this section are made formally precise in subsequent sections.

6.2 Constrained Input/Output Logic

We briefly recall the basics of *Constrained Input/Output logic*, the systems for which we provide argumentative characterizations. The formalism was developed by Makinson and van der Torre (2001) and is particularly suitable for normative reasoning (Parent and van der Torre, 2018b). Its central feature is the employment of syntactic objects of the form (φ, ψ), called *norms*.[6] I/O logics are construed over the *non*-labelled propositional language $\mathcal{L}$ of Definition 6.1 and a propositional base logic L. We use Greek capital letters $\Delta, \Gamma, \ldots$ for finite sets of $\mathcal{L}$-formulae and write $\bigwedge \Delta$ to denote the conjunction of elements of Δ. To obtain results as general as possible, we define a large class of underlying base logics for which we provide the envisioned results in this chapter. Definition 6.2 stipulates the properties these base logics must satisfy. This class includes but is not limited to classical and intuitionistic logic.

Definition 6.2 (Base Logic L)**.** *Let* $\mathsf{L} = \langle \mathcal{L}, \vdash \rangle$ *be a* propositional logic*, where* $\mathcal{L}$ *is the language from Definition 6.1 and* $\vdash_{\mathsf{L}}$ *denotes the consequence relation on* $\wp(\mathcal{L}) \times \mathcal{L}$ *(henceforth, we omit the subscript* L*). Let* $\mathcal{S} \subseteq \mathcal{L}$*,* $\vdash$ *satisfies the following properties:*

- reflexivity*: if* $\varphi \in \mathcal{S}$*, then* $\mathcal{S} \vdash \varphi$*;*
- monotonicity*: if* $\mathcal{S}' \vdash \varphi$ *and* $\mathcal{S}' \subseteq \mathcal{S}$*, then* $\mathcal{S} \vdash \varphi$*;*
- transitivity*: if* $\mathcal{S} \vdash \varphi$ *and* $\mathcal{S}', \varphi \vdash \psi$ *then* $\mathcal{S}, \mathcal{S}' \vdash \psi$*;*
- non-triviality*:* $\mathcal{S} \nvdash \varphi$ *for some* $\mathcal{S} \neq \emptyset$ *and* φ*;*
- structurality*: if* $\mathcal{S} \vdash \varphi$*, then* $\{\theta(\psi) \mid \psi \in \mathcal{S}\} \vdash \theta(\varphi)$ *for every substitution* θ*;*
- compactness*: if* $\mathcal{S} \vdash \varphi$ *then* $\Gamma \vdash \varphi$ *for some finite* $\Gamma \subseteq \mathcal{S}$*.*

Furthermore, we assume that conjunction $\wedge$ *and disjunction* $\vee$ *are distributive, i.e.,* $\vdash (\varphi \wedge \psi_1) \vee (\varphi \wedge \psi_2) \equiv (\varphi \wedge (\psi_1 \vee \psi_2))$ *and* $\vdash (\varphi \vee \psi_1) \wedge (\varphi \vee \psi_2) \equiv (\varphi \vee (\psi_1 \wedge \psi_2))$*.*

Constrained I/O logics work with knowledge bases as defined below.

Definition 6.3 (Knowledge Base $\mathcal{K}$)**.** *Let* L *be the underlying base logic. Let* $\mathcal{K} = \langle \mathcal{F}, \mathcal{N}, \mathcal{C} \rangle$ *be a* knowledge base*, where* $\mathcal{F} \subseteq \mathcal{L}$ *is the* factual *input,* $\mathcal{N} \subseteq \mathcal{L} \times \mathcal{L}$ *a* normative code*, and* $\mathcal{C} \subseteq \mathcal{L}$ *a set of* constraints *containing the formulae with which output must be consistent. We assume* $\mathcal{F}$ *and* $\mathcal{C}$ *to be consistent, i.e.,* $\mathcal{F} \nvdash \bot$ *and* $\mathcal{C} \nvdash \bot$*. The sets* $\mathcal{F}$*,* $\mathcal{N}$*, and* $\mathcal{C}$ *may be countably infinite.*

[6] The Input/Output formalism was applied by Bochman (2021) to causal and doxastic reasoning, where a pair (φ, ψ) is interpreted as "φ causes ψ", respectively "believing φ leads to believing ψ". Ciabattoni et al. (2021) extend the I/O formalism to model legal reasoning in the spirit of Kelsen's theory of norms. More general, one can read (φ, ψ) as a generator, i.e., "input (premises) φ generates output (conclusion) ψ". We discuss other applications in Section 6.8.

$$\frac{}{(\top,\top)}\,\mathrm{T} \qquad\qquad \frac{}{(\varphi,\varphi)}\,\mathrm{ID}$$

$$\frac{(\varphi,\psi) \qquad \psi\vdash\gamma}{(\varphi,\gamma)}\,\mathrm{WO} \qquad \frac{(\varphi,\psi) \qquad (\varphi,\gamma)}{(\varphi,\psi\wedge\gamma)}\,\mathrm{AND} \qquad \frac{(\varphi,\psi) \qquad \gamma\vdash\varphi}{(\gamma,\psi)}\,\mathrm{SI}$$

$$\frac{(\varphi,\psi) \qquad (\varphi\wedge\psi,\gamma)}{(\varphi,\gamma)}\,\mathrm{CT} \qquad \frac{(\varphi,\psi) \qquad (\gamma,\psi)}{(\varphi\vee\gamma,\psi)}\,\mathrm{OR}$$

Figure 6.3: Rules for constructing $\mathsf{deriv}_{\mathcal{R},\mathsf{L}}$ proof systems. The topmost level contains initial rules. The minimal set of deriv-rules is $\{\mathrm{WO},\mathrm{AND},\mathrm{SI}\}$.

Proof theory for the I/O formalism was introduced by Makinson and van der Torre (2000) for a class of *monotonic* I/O logics. The resulting systems are referred to as "deriv" and contain inference rules that derive I/O pairs from other I/O pairs. The deriv-rules considered in this chapter are those developed by Makinson and van der Torre (2000) and presented in Table 6.3. We refer to the work of Parent and van der Torre (2018b) for a discussion of other deriv rules.

Definition 6.4 (unconstrained I/O proof systems)**.** *Let* $\mathsf{deriv}_{\mathcal{R},\mathsf{L}}$ *be a proof system, with* $\mathcal{R}$ *a set of rules from Table 6.3. Let* L *be the base logic, and let* $\mathcal{N}\subseteq\mathcal{L}^n$*. A* derivation *of* $(\varphi,\psi)\in\mathsf{deriv}_{\mathcal{R},\mathsf{L}}(\mathcal{N})$ *is a* tree *of rule-applications of* $\mathcal{R}$ *where the leaves are either members of* $\mathcal{N}$ *or instances of T and ID (provided T, ID* $\in\mathcal{R}$*), and the root is* (φ,ψ)*.*

We say ψ *is* obligatory *(detached) under* $\mathcal{N}$ *and* $\mathcal{F}$ *if* $(\varphi,\psi)\in\mathsf{deriv}_{\mathcal{R},\mathsf{L}}(\mathcal{N})$ *with* $\mathcal{F}\vdash\varphi$*. We write* $\psi\in\mathsf{deriv}_{\mathcal{R},\mathsf{L}}(\Delta,\mathcal{N})$ *if* $(\bigwedge\Delta,\varphi)\in\mathsf{deriv}_{\mathcal{R},\mathsf{L}}(\mathcal{N})$*.*

Paradigmatic I/O logics are characterized by the sets of rules $\mathcal{R}_1=\{\mathrm{T},\mathrm{WO},\mathrm{SI},\mathrm{AND}\}$, $\mathcal{R}_2=\{\mathrm{OR}\}\cup\mathcal{R}_1$, $\mathcal{R}_3=\mathcal{R}_1\cup\{\mathrm{CT}\}$, and $\mathcal{R}_4=\mathcal{R}_2\cup\mathcal{R}_3$. The system $\mathcal{R}_1$ represents a *single deontic detachment* procedure which allows for weakening of the output (WO), combining output (AND), and strengthening of the input (SI). All propositional tautologies are among the output (T). System $\mathcal{R}_2$ extends $\mathcal{R}_1$ with *reasoning by cases* (OR), i.e., if both φ and γ generate output ψ, then $\varphi\vee\gamma$ generates ψ too. System $\mathcal{R}_3$ extends $\mathcal{R}_1$ with reusability (CT) allowing for iterations of *successive deontic detachment* (cf. chaining reasons in Example 6.4).[7] Last, $\mathcal{R}_4$ combines reasoning by cases $\mathcal{R}_2$ and successive detachment $\mathcal{R}_3$. The above systems may be closed under *throughput* (ID), i.e., input is 'put through' as output. Throughput indicates that facts are considered as part of the obligations in the output.[8] We write $\mathcal{R}_i^+=\mathcal{R}_i\cup\{\mathrm{ID}\}$ for $i\in\{1,2,3,4\}$. The resulting eight systems are sound and complete with respect to their semantic characterizations (Parent and van der Torre, 2018b). We omit the semantics here.

[7] CT stands for cumulative transitivity.

[8] A more intuitive application of ID would be in the context of doxastic reasoning.

Remark 6.1 (Free-Floaters)**.** *Consider a knowledge base $\mathcal{K} = \langle \mathcal{F}, \mathcal{N}, \mathcal{C} \rangle$. A* free-floater *is a norm $(\varphi, \psi) \in \mathcal{N}$ that cannot be (indirectly) triggered, in any way, by the context $\mathcal{F}$. To illustrate, let $\mathcal{F} = \emptyset$ and $\mathcal{N} = \{(p,q), (\top, r)\}$. It can be straightforwardly observed that (p,q) is a free-floater for all of the* $\mathsf{deriv}_{\mathcal{R},\mathsf{L}}$ *systems for which $OR \notin \mathcal{R}$ (*L *is consistent and non-trivial). Formally, we say (φ, ψ) is a free-floater with respect to $\mathcal{K} = \langle \mathcal{F}, \mathcal{N}, \mathcal{C} \rangle$ and* $\mathsf{deriv}_{\mathcal{R},\mathsf{L}}$ *whenever for all derivations with root (θ, γ) containing (φ, ψ) as a leaf, $\mathcal{F} \nvdash \theta$. Since free-floaters cannot be triggered at all, they do not influence the detachable obligations of any of the eight Input/Output operations. In what follows, we only consider knowledge bases void of free-floaters.*

Remark 6.2 (Relevance)**.** *Through the application of rules, a derivation contains only norms used for deriving the norm at its root. However, this does not guarantee that all norms in the derivation are relevant for deriving the derivation's conclusion. Let $\mathcal{F} = \{p\}$ and $\mathcal{N} = \{(p,q), (p,r)\}$ be the knowledge base, the following two example* deriv*-derivations demonstrate that* deriv *cannot assure relevance in derivations (cf. Section 6.1).*

$$\dfrac{}{(p,q)} \qquad\qquad \dfrac{\dfrac{\dfrac{}{(p,q)} \quad \dfrac{}{(p,r)}}{(p, q \wedge r)}\,\mathit{AND}}{(p,q)}\,\mathit{WO}$$

Both derivations show that q is obligatory (detached) under $\mathcal{N}$ and $\mathcal{F}$. The derivation on the left derives q from the norm set $\mathcal{N}' = \{(p,q)\}$, i.e., $q \in$ $\mathsf{deriv}_{\mathcal{R},\mathsf{L}}(\mathcal{F}, \mathcal{N}')$*. The derivation on the right, however, also derives q but from the norm set $\mathcal{N} = \{(p,q), (p,r)\}$, i.e., $q \in$* $\mathsf{deriv}_{\mathcal{R},\mathsf{L}}(\mathcal{F}, \mathcal{N})$*. Clearly, $\mathcal{N}'$ is more relevant than $\mathcal{N}$ in explaining the obligation q. In fact, $\mathcal{N}'$ is the only relevant set explaining q. The approach we presented in Section 6.7 does preserve relevance in explaining obligations.*

The above deriv systems are still monotonic. Scenarios, such as those presented in Examples 6.1 and 6.3, can be effectively addressed using methods of nonmonotonic reasoning. To see that unconstrained monotonic deriv leads to problems, reconsider Example 6.1 with $\mathcal{F} = \{\neg h\}$ and $\mathcal{N} = \{(\top, h), (h, t), (\neg h, \neg t)\}$. The following derivation shows that the example is inconsistent in a monotonic I/O setting.

$$\dfrac{\neg t \wedge t \vdash \bot \qquad \dfrac{\dfrac{\dfrac{\dfrac{}{(\top, h)} \quad \dfrac{\dfrac{}{(h,t)} \quad \top \wedge h \vdash h}{(\top \wedge h, t)}\,\mathrm{SI}}{(\top, t)}\,\mathrm{CT} \quad \top \wedge \neg h \vdash \top}{(\top \wedge \neg h, t)}\,\mathrm{SI} \quad \dfrac{\dfrac{}{(\neg h, \neg t)} \quad \top \wedge \neg h \vdash \neg h}{(\top \wedge \neg h, \neg t)}\,\mathrm{SI}}{(\top \wedge \neg h, \neg t \wedge t)}\,\mathrm{AND}}{(\top \wedge \neg h, \bot)}\,\mathrm{WO}$$

That is, $\bot \in$ $\mathsf{deriv}_{\mathcal{R},\mathsf{L}}(\mathcal{F}, \mathcal{N})$ (with CT $\in \mathcal{R}$). The derivation shows that given the context $\mathcal{F}$, we can derive both the obligation to tell and the obligation not to tell. Instead, we want to derive Billy' obligation given that she already violated the norm $(\top, h) \in \mathcal{N}$, i.e., given the fact $\neg h \in \mathcal{F}$. For this, we need defeasible detachment.

The solution developed by Makinson and van der Torre (2001) is to put constraints on I/O reasoning. The resulting nonmonotonic systems are called *constrained I/O logics* (Makinson and van der Torre, 2001).[9] The use of constraints may yield various maximal sets of norms $\mathcal{N}' \subseteq \mathcal{N}$ whose detachable output is consistent with the constraints $\mathcal{C}$. If the output is required to be consistent *per se*, it suffices to let $\mathcal{C}$ be empty, i.e., $\mathcal{C} = \emptyset$. If the output is to be consistent with the input, we take the factual context to be a subset of the constraints, i.e., $\mathcal{F} \subseteq \mathcal{C}$ (e.g., Example 6.2). Constrained I/O logics draw inferences from such maximally consistent norm sets. As is common to nonmonotonic reasoning in general, there are two main approaches to drawing conclusions from such sets. *Skeptic* inference allows us only to infer what is common to all of these sets. *Credulous* inference considers each set as an acceptable solution, and we may infer a formula whenever it belongs to one such set. Let us make this formally precise.

Definition 6.5 (Constrained I/O Logics)**.** *Let* $\mathsf{deriv}_{\mathcal{R},\mathsf{L}}$ *be a system from Figure 6.3 and let* $\mathcal{K} = \langle \mathcal{F}, \mathcal{N}, \mathcal{C} \rangle$ *be a knowledge base void of free-floaters (Remark 6.1). The set of* maximally consistent families *of* $\mathcal{N}$ *(*maxfam*) is defined as:*

- $\mathsf{maxfam}_{\mathcal{R},\mathsf{L}}(\mathcal{K})$ *is the set of* max-elements *of* $\{\mathcal{N}' \subseteq \mathcal{N} \mid$ *for all* $(\varphi, \psi) \in \mathsf{deriv}_{\mathcal{R},\mathsf{L}}(\mathcal{N}')$, *if* $\mathcal{F} \vdash \varphi$, *then* $\mathcal{C}, \psi \nvdash \bot\}$.

We define skeptic ($\mathrel{|\!\sim}^{s}$)*, respectively credulous* ($\mathrel{|\!\sim}^{c}$) *nonmonotonic inference for* constrained I/O logic *as follows:*

- $\mathcal{K} \mathrel{|\!\sim}^{s}_{\mathcal{R},\mathsf{L}} \varphi$ *iff for each* $\mathcal{N}' \in \mathsf{maxfam}_{\mathcal{R},\mathsf{L}}(\mathcal{K})$*, there is a* $(\psi, \varphi) \in \mathsf{deriv}_{\mathcal{R},\mathsf{L}}(\mathcal{N}')$ *with* $\mathcal{F} \vdash \psi$*;*
- $\mathcal{K} \mathrel{|\!\sim}^{c}_{\mathcal{R},\mathsf{L}} \varphi$ *iff for some* $\mathcal{N}' \in \mathsf{maxfam}_{\mathcal{R},\mathsf{L}}(\mathcal{K})$*, there is a* $(\psi, \varphi) \in \mathsf{deriv}_{\mathcal{R},\mathsf{L}}(\mathcal{N}')$ *with* $\mathcal{F} \vdash \psi$*.*

Example 6.4 (Example 6.1 cont.)**.** *Reconsider the Chisholm scenario, where* $\mathcal{F} = \{\neg h\}$*, and* $\mathcal{N} = \{(\top, h), (h, t), (\neg h, \neg t)\}$*. Let the I/O system be defined by* $\mathcal{R}_3$ *and a classical base logic* L*. For the empty constraint set* $\mathcal{C} = \emptyset$*, we have* $\mathsf{maxfam}_{\mathcal{R}_3,\mathsf{L}}(\mathcal{F}, \mathcal{N}, \mathcal{C}) = \{\{(\top, h), (h, t)\}, \{(\neg h, \neg t), (\top, h)\}, \{(\neg h, \neg t), (h, t)\}\}$*. In other words, we have both* $\mathcal{K} \not\mathrel{|\!\sim}^{s}_{\mathcal{R}_3,\mathsf{L}} h \wedge t$ *and* $\mathcal{K} \not\mathrel{|\!\sim}^{s}_{\mathcal{R}_3,\mathsf{L}} \neg t$*. Still, we conclude* $\mathcal{K} \mathrel{|\!\sim}^{c}_{\mathcal{R}_3,\mathsf{L}} h \wedge t$ *and* $\mathcal{K} \mathrel{|\!\sim}^{c}_{\mathcal{R}_3,\mathsf{L}} h \wedge t$*. For instance, we derive* $(\top, h \wedge t) \in \mathsf{deriv}_{\mathcal{R}_3,\mathsf{L}}(\{(\top, h), (h, t)\})$ *as follows:*

$$\dfrac{(\top,h) \qquad \dfrac{(\top,h) \qquad \dfrac{(h,t) \qquad \top\wedge h\vdash h}{(\top\wedge h,t)}\,SI}{(\top,t)}\,CT}{(\top,h\wedge t)}\,AND$$

[9] We stress that these logics were only introduced semantically.

Furthermore, since $\mathcal{F} \vdash \top$ *and* $\mathcal{C}, h \wedge t \nvdash \bot$*, we have* $\{(\top, h), (h, t)\} \in \mathsf{maxfam}_{\mathcal{R}_3,\mathsf{L}}(\mathcal{F}, \mathcal{N}, \mathcal{C})$ *(it can be straightforwardly observed that this set is also maximal).*

However, once we set the constraints to Billy's violation, i.e., $\mathcal{C}' = \mathcal{F}$*, we obtain a singleton set* $\mathsf{maxfam}_{\mathcal{R}_3,\mathsf{L}}(\mathcal{F}, \mathcal{N}, \mathcal{C}) = \{\mathcal{N}'\}$*, with* $\mathcal{N}' = \{(\neg h, \neg t), (h, t)\}$*. Intuitively, only* $\mathcal{N}'$ *remains because* $\mathcal{C}', h \vdash \bot$ *and* $\mathcal{C}', \neg t \nvdash \bot$*. (In addition, observe that* $(h, t) \in \mathcal{N}'$ *cannot be triggered by* $\mathcal{F}$*.) Given* $\mathcal{C}'$*, Billy is therefore obliged not to tell her neighbors she is coming to help, i.e.,* $\mathcal{K} \mid\!\sim^{s}_{\mathcal{R}_3,\mathsf{L}} \neg t$*. What is more, Billy is not obliged to help in this contrary-to-duty scenario, i.e.,* $\mathcal{K} \not\mid\!\sim^{s}_{\mathcal{R}_3,\mathsf{L}} h$*.*

Before introducing Deontic Argumentation Calculi in the next section, we stress the following three points: First, deriv does not guarantee that all norms used in a derivation are strictly *relevant* for the derivation's conclusion (i.e., Remark 6.2). Second, maxfam sets (of arbitrary size) do not provide formal ways of pinpointing the reasons why certain norms are *inapplicable*. For instance, the set $\mathsf{maxfam}_{\mathcal{R}_3,\mathsf{L}}(\mathcal{F}, \mathcal{N}, \mathcal{C}) = \{\{(\neg h, \neg t), (h, t)\}\}$ in Example 6.4 does not explain why $(\top, h)$ is inapplicable given $\mathcal{C} = \mathcal{F}$. Third, deriv is not suitable for generating transparent arguments. For example, as a certificate the derivation in Example 6.4 may justify *that* $(\top, h \wedge t)$ is derivable, its conclusion does not explain *why* $h \wedge t$ is obligatory (cf. page 213). Our calculi address all three challenges.

6.3 Deontic Argumentation Calculi

Our first step towards realizing more transparent I/O arguments, is to label propositional formulae as facts $\mathcal{L}^f$, obligations $\mathcal{L}^o$, and constraints $\mathcal{L}^c$. Second, we allow for Boolean operations over the more complex meta-logical objects (φ, ψ) denoting norms. Operations over these higher-order syntactic objects enable undercuts that explain why certain norms should (not) be applied. For the present chapter, it suffices to consider negation only. A formula $\neg(\varphi, \psi)$ is interpreted as "the norm (φ, ψ) is inapplicable". The full labelled I/O language is given below.

Definition 6.6 (The Labelled I/O Language $\mathcal{L}^{io}$)**.** *Let* $\mathcal{L}^i$ *with* $i \in \{f, o, c, n\}$ *be as defined in Definition 6.1. The* language of norms *is defined as* $\mathcal{L}^n \cup \overline{\mathcal{L}^n}$*, where* $\overline{\mathcal{L}^n} = \{\neg(\varphi, \psi) \mid (\varphi, \psi) \in \mathcal{L}^n\}$ *is the language expressing the inapplicability of norms. Let* $\mathcal{L}^{io} = \mathcal{L}^f \cup \mathcal{L}^o \cup \mathcal{L}^c \cup \mathcal{L}^n \cup \overline{\mathcal{L}^n}$ *be the* labelled I/O language.

We write φ *for an arbitrary formula of* $\mathcal{L}^{io}$ *and write* Δ^i *to denote the labelled set of formulae* $\Delta^i \subseteq \mathcal{L}^i$ *with* $i \in \{f, o, c, n\}$*.*

Our aim is to develop a large class of sequent-style calculi characterizing various I/O logics (Objective 2). For this purpose, we introduce Deontic Argumentation Calculi, DAC for short. These calculi are *sequent-style* proof systems (Gentzen, 1934). The main syntactic object of a sequent-style calculus is that of a *sequent*, i.e., $\Delta \Rightarrow \Gamma$ where Δ and Γ are sets of formulae. Sequent calculi are proof systems characterized by sets of *rules* (cf. Hilbert-style proof systems, which take axioms and a few inference rules as their central components). Such rules lay down the conditions under which sequents may be derived from other sequents. By using rules, sequent systems generate proofs as trees: The leaves of a tree are sequents that are either trivial logical truths or assumptions, branches are the result of rule applications, and the root of the tree is the proof's concluding sequent (Negri et al., 2008).

Remark 6.3. *We refer to $\Delta \Rightarrow \Gamma$ as an* argument*, where Δ denotes the reasons for Γ; cf. Section 6.1 and (Arieli and Straßer, 2015). In the remainder of this chapter, we interpret Δ on the lhs of an argument $\Delta \Rightarrow \Gamma$ as a* regular finite set*. The use of regular sets instead of multi-sets accommodates a higher modularity with respect to the underlying base logic L. Moreover, we only consider arguments with* single-conclusions*, that is, whose rhs is either a formula or the empty-set. In the sequel, we assume that a set Γ on the rhs of an argument $\Delta \Rightarrow \Gamma$ is restricted to at most one formula.*

For each base logic of Definition 6.2 we assume a corresponding sound and complete *sequent calculus* LC. In Definition 6.7, we stipulate some minimal properties that the sequent calculus LC must satisfy.

Definition 6.7. *Let L be a base logic of Definition 6.2 over the language $\mathcal{L}$ of Definition 6.1. Let LC be a* sequent calculus *such that LC is sound and complete for L. That is, for each $\Delta \subseteq \mathcal{L}$ and $\varphi \in \mathcal{L}$, $\Delta \vdash \varphi$ iff the sequent $\Delta \Rightarrow \varphi$ is LC-derivable. Furthermore, we assume admissibility of the following logical rules in LC:*

$$\dfrac{\varphi \Rightarrow \psi_1 \qquad \varphi \Rightarrow \psi_2}{\varphi \Rightarrow \psi_1 \wedge \psi_2}\ R\wedge 1 \qquad \dfrac{\varphi_1, \varphi_2 \Rightarrow \psi}{\varphi_1 \wedge \varphi_2 \Rightarrow \psi}\ R\wedge 2 \qquad \dfrac{\varphi_1 \Rightarrow \psi \qquad \varphi_2 \Rightarrow \psi}{\varphi_1 \vee \varphi_2 \Rightarrow \psi}\ R\vee 2$$

$$\dfrac{\varphi \Rightarrow \psi_1}{\varphi \Rightarrow \psi_1 \vee \psi_2}\ R\vee 1 \qquad \dfrac{\varphi \Rightarrow \psi}{\varphi, \neg\psi \Rightarrow}\ R\neg 1 \qquad \dfrac{\varphi, \psi \Rightarrow}{\varphi \Rightarrow \neg\psi}\ R\neg 2$$

$$\dfrac{}{\bot \Rightarrow \psi}\ \bot \qquad \dfrac{\Delta \Rightarrow}{\Delta \Rightarrow \bot}\ R\bot$$

The top sequents of a rule denote the rule's premises (or conditions). The bottom-sequent expresses the conclusion of a rule. Double lines in a rule indicate the invertibility of the rule, i.e., derivability in both directions.[10] *We refer to the work of Negri et al. (2008) for an extensive introduction to sequent-style proof systems.*

Definition 6.8 (Deontic Argumentation Calculi)**.** *Let LC be the underlying base calculus. We say $\Delta \Rightarrow \Gamma$ is a* DAC-sequent *whenever $\Delta \subseteq \mathcal{L}^{io}$ is a finite set, and $\Gamma \subseteq \mathcal{L}^{io}$ is a set restricted to at most one formula.*

The minimal system, referred to as $\mathsf{DAC}_{\emptyset}$, contains the rules **Ax**, **Detach**, **R-C**, **R-N**, *and* **Cut** *from Figure 6.4. The* calculus $\mathsf{DAC}_{\mathcal{S}}$ *extends $\mathsf{DAC}_{\emptyset}$ with the set of rules*

[10] Notice that the inverted version of an invertible rule with two premises corresponds to two rules. For instance, given $\varphi \Rightarrow \psi_1 \wedge \psi_2$, the inverted rules of $R \wedge 1$ derive $\varphi \Rightarrow \psi_1$, respectively $\varphi \Rightarrow \psi_2$.

$\mathcal{S} \subseteq \{\textbf{Taut},\textbf{TP},\textbf{L-OR},\textbf{L-CT}\}$. *This leads to a total of 16* DAC*-systems. We write* $\mathcal{S}^+$ *when* **TP** $\in \mathcal{S}$ *and* $\mathcal{S}^-$ *when* **TP** $\notin \mathcal{S}$.

A $\mathsf{DAC}_{\mathcal{S}}$-derivation *of* $\Delta \Rightarrow \Gamma$ *is a* tree *whose leaves are initial sequents of* $\mathsf{DAC}_{\mathcal{S}}$*, whose root is* $\Delta \Rightarrow \Gamma$*, and whose rule-applications are instances of the rules of* $\mathsf{DAC}_{\mathcal{S}}$*. We write* $\vdash_{\mathcal{S}} \Delta \Rightarrow \Gamma$ *if* $\Delta \Rightarrow \Gamma$ *is* $\mathsf{DAC}_{\mathcal{S}}$*-derivable. We write* $\vdash^n_{\mathcal{S}} \Delta \Rightarrow \Gamma$ *if* $\Delta \Rightarrow \Gamma$ *is* $\mathsf{DAC}_{\mathcal{S}}$*-derivable in at most n steps (a step is defined by a rule application).*

Definition 6.8 stipulates that all propositional formulae occurring in a DAC-sequent are labelled f, o, or c. Let LC be an adequate sequent calculus for the base logic L, then, intuitively, DAC takes *labelled* versions of any LC-derivable $\Delta \Rightarrow \Gamma$ as an initial sequent (i.e., $\Delta^i \Rightarrow \Gamma^i$ for each $i \in \{f, o, c\}$) and contains logical- and structural rules for transforming labelled formulae of $\mathcal{L}^{io}$ (Figure 6.4). Since $\mathsf{DAC}_{\mathcal{S}}$ takes labelled LC-derivable sequents as initial sequents, the rules of LC are not part of $\mathsf{DAC}_{\mathcal{S}}$. Still, LC rules can be straightforwardly shown admissible in DAC due to the presence of **Cut**. For instance, that the rule R∧1 in Definition 6.7 is admissible in DAC is shown by the following derivation (where $\Delta, \Gamma \subseteq \mathcal{L}^{io}$):

$$\dfrac{\Delta \Rightarrow \varphi^i \qquad \dfrac{\Gamma \Rightarrow \psi^i \qquad \varphi^i, \psi^i \Rightarrow (\varphi \wedge \psi)^i}{\varphi^i, \Gamma \Rightarrow (\varphi \wedge \psi)^i}\,\textbf{Cut}}{\Delta, \Gamma \Rightarrow (\varphi \wedge \psi)^i}\,\textbf{Cut}$$

The rule **Taut** ensures that all propositional tautologies are considered as output. The rule **Detach** is an initial explanatory argument stating that the fact φ and the norm (φ, ψ) are reasons for the obligation ψ. Instead of deriving pairs from other pairs (as in deriv), we keep norms as primitive reasons from a given normative code $\mathcal{N}$ and only modify facts, obligations, and constraints. This gives us some explanatory advantages (see **R-C** and **R-N** below). The rule **TP** corresponds to throughput. The rule **L-CT** corresponds to successive detachment, expressing that a norm may likewise be triggered by the output of some other norm (cf. Example 6.5). **L-OR** reflects reasoning by cases over input. The side-condition on **L-OR** is dropped whenever **TP** $\in \mathcal{S}$. Namely, since **TP** allows us to conclude obligations from facts without the intermediate use of norms, **L-OR** no longer requires the presence of norms in the premises of the rule when **TP** is part of the calculus. **Cut** suffices as the only structural rule. The absence of the structural rule *weakening* in DAC (i.e., $\Delta \Rightarrow \Gamma$ implies $\varphi, \Delta \Rightarrow \Gamma$, for $\varphi \in \mathcal{L}^{io}$) is deliberate because we desire *relevance* in constructing arguments. Namely, only those norms relevant for explaining a particular obligation or defeat are present in the argument. Although the absence of weakening is necessary, in Section 6.7 we show that it is not sufficient, and extend the calculi with relevance rules.

More interesting for explainability are the rules **R-C** and **R-N**. Concerning **R-C**, think of a sequent with an empty right-hand side as an argument expressing inconsistent reasons. For instance, an argument $\varphi^f, (\varphi, \psi), (\neg\psi)^c \Rightarrow$ explains that the fact φ and the norm (φ, ψ) (which are reasons for ψ) are inconsistent whenever the output must

$$\dfrac{}{\Delta^i \Rightarrow \Gamma^i}\,\textbf{Ax}\text{, for each } \mathsf{LC}\text{-derivable } \Delta^{\downarrow} \Rightarrow \Gamma^{\downarrow}\text{, with } i \in \{f, o, c\} \text{ and } \Delta^{\downarrow}, \Gamma^{\downarrow} \subseteq \mathcal{L}$$

$$\dfrac{}{\Rightarrow (\top, \top)}\,\textbf{Taut} \qquad \dfrac{}{\varphi^f, (\varphi, \psi) \Rightarrow \psi^o}\,\textbf{Detach} \qquad \dfrac{}{\varphi^f \Rightarrow \varphi^o}\,\textbf{TP}$$

$$\dfrac{\Delta \Rightarrow \varphi^o}{\Delta, (\neg\varphi)^c \Rightarrow}\,\textbf{R-C} \qquad \dfrac{\Delta, (\varphi, \psi) \Rightarrow}{\Delta \Rightarrow \neg(\varphi, \psi)}\,\textbf{R-N} \qquad \dfrac{\varphi^f, \Delta \Rightarrow \Gamma}{\varphi^o, \Delta \Rightarrow \Gamma}\,\textbf{L-CT}^a$$

$$\dfrac{\Delta, \varphi^f \Rightarrow \Gamma \qquad \Delta', \psi^f \Rightarrow \Gamma}{\Delta, \Delta', (\varphi \vee \psi)^f \Rightarrow \Gamma}\,\textbf{L-OR}^b \qquad \dfrac{\Delta \Rightarrow \varphi \qquad \varphi, \Delta' \Rightarrow \Gamma}{\Delta, \Delta' \Rightarrow \Gamma}\,\textbf{Cut}^c$$

Figure 6.4: Rules for constructing $\mathsf{DAC}_{\mathcal{S}}$ (Definition 6.8). The upper level contains initial sequents, and the lower level logical and structural rules. For **Ax**, let $\Delta^{\downarrow} := \{\, \varphi \mid \varphi^i \in \Delta^i \,\}$ denote the set Δ^i stripped from its labels. Side-condition (a) on **L-CT** denotes $\Delta \cap \mathcal{L}^n \neq \emptyset$; (b) on **L-OR** denotes $\Delta \cap \mathcal{L}^n \neq \emptyset$ and $\Delta' \cap \mathcal{L}^n \neq \emptyset$, and is only imposed when $\textbf{TP} \notin \mathcal{S}$; (c) on **Cut** stipulates that $\varphi \in \mathcal{L}^{io}$.

be consistent with $\neg\psi$. Moreover, whenever such an argument expresses inconsistent reasons, we know at least one of its norms is inapplicable. The rule **R-N** expresses this: from $\varphi^f, (\varphi, \psi), (\neg\psi)^c \Rightarrow$ we obtain the defeating argument $\varphi^f, (\neg\psi)^c \Rightarrow \neg(\varphi, \psi)$. Hence, φ^f and $(\neg\psi)^c$ are reasons for the *inapplicability* of the norm (φ, ψ). $\mathsf{DAC}_{\mathcal{S}}$ sequents will be the building blocks for the desired argumentative characterizations (Section 6.5).

Example 6.5 (Example 6.1 cont.)**.** *The* DAC*-argument* d *in Figure 6.1, concluding that Billy should tell her neighbors she is coming to help, is derived through chaining* $(\top, h)$ *and* (h, t)*. The* $\mathsf{DAC}_{\mathcal{S}}$*-derivation below on the left demonstrates this (where* $\textbf{L-CT} \in \mathcal{S}$*).*

$$\dfrac{\dfrac{}{\top^f, (\top, h) \Rightarrow h^o}\,\textbf{Detach} \qquad \dfrac{\dfrac{}{h^f, (h, t) \Rightarrow t^o}\,\textbf{Detach}}{h^o, (h, t) \Rightarrow t^o}\,\textbf{L-CT}}{\top^f, (\top, h), (h, t) \Rightarrow t^o}\,\textbf{Cut} \qquad\qquad \dfrac{\dfrac{\dfrac{}{\top^f, (\top, h) \Rightarrow h^o}\,\textbf{Detach}}{\top^f, (\neg h)^c, (\top, h) \Rightarrow}\,\textbf{R-C}}{\top^f, (\neg h)^c \Rightarrow \neg(\top, h)}\,\textbf{R-N}$$

Given the constraint $\mathcal{C}' = \{\neg h^c\}$*, the question "why should Billy not tell she is coming to help,* despite *argument* d*?" is answered by argument* b *in Figure 6.1. A* $\mathsf{DAC}_{\mathcal{S}}$*-derivation of argument* b *is provided above on the right. The fact* $\top^f$ *can be omitted by an application of* **Cut** *with the sequent* $\Rightarrow \top^f$*.*

Example 6.6 (Example 6.3 cont.)**.** *In the dilemma, Joan cannot both return the hammer and prevent harm from being done. So, we find* $(\top, r)$ *applicable if and only if* $(\top, p)$ *is inapplicable. This is expressed by arguments* e *and* f*. The following* $\mathsf{DAC}_{\mathcal{S}}$*-derivation shows the derivability of argument* e*:*

$$\dfrac{\dfrac{\dfrac{\dfrac{}{\top^f,(\top,r)\Rightarrow r^o}\,\textbf{Detach}\qquad\dfrac{}{\top^f,(\top,p)\Rightarrow p^o}\,\textbf{Detach}}{\top^f,(\top,r),(\top,p)\Rightarrow(r\wedge p)^o}\,\mathrm{R}\wedge 1}{\top^f,(\top,r),(\top,p),\neg(r\wedge p)^c\Rightarrow}\,\textbf{R-C}}{\top^f,(\top,p),\neg(r\wedge p)^c\Rightarrow\neg(\top,r)}\,\textbf{R-N}$$

The LC*-rule R∧1 used in the above derivation is* DAC*-admissible (see page 226). We can apply the* **R-N** *rule to* $(\top,p)$ *in the above derivation to obtain argument* f*.*

6.4 Soundness and Completeness, Part 1

In what follows, we demonstrate the first of our two soundness and completeness proofs: the correspondence between the I/O proof system deriv and our proof system DAC. We start by proving general lemmas concerning the inference relation $\vdash_{\mathcal{S}}$.

6.4.1 Some Technical Lemmas Concerning DAC

First, we prove some facts about derivability in the base logic LC, which are used in the proofs below (without further reference).

Observation 6.1. *Let* LC *be the underlying base logic, the following hold:*

1. $\vdash_{\mathsf{LC}} \psi\Rightarrow\neg\neg\psi$
2. *If* $\vdash_{\mathsf{LC}} \varphi\Rightarrow\neg\bigwedge\Delta$ *then* $\vdash_{\mathsf{LC}} \varphi,\Delta\Rightarrow$
3. *If* $\vdash_{\mathsf{LC}} \varphi\Rightarrow\neg\psi_1$ *then* $\vdash_{\mathsf{LC}} \varphi\Rightarrow\neg(\psi_1\wedge\psi_2)$
4. *If* $\varphi_1\Rightarrow\neg\bigwedge\Delta_1$ *and* $\vdash_{\mathsf{LC}} \varphi_2\Rightarrow\neg\bigwedge\Delta_2$ *then* $\vdash_{\mathsf{LC}} \varphi_1\vee\varphi_2\Rightarrow\neg\bigwedge\Delta_1\cup\Delta_2$

Proof. We consider each item consecutively:

Ad 1. The following derivation suffices:

$$\dfrac{\dfrac{\dfrac{}{\vdash_{\mathsf{LC}}\psi\Rightarrow\psi}\,\text{by the reflexivity of }\vdash}{\psi,\neg\psi\Rightarrow}\,\mathrm{R}\neg 1}{\psi\Rightarrow\neg\neg\psi}\,\mathrm{R}\neg 2$$

Ad 2. The following derivation suffices:

$$\dfrac{\dfrac{\dfrac{}{\bigwedge\Delta\Rightarrow\neg\neg\bigwedge\Delta}\,\text{Item 1}}{\Delta\Rightarrow\neg\neg\bigwedge\Delta}\,\mathrm{R}\wedge 2\qquad\dfrac{\dfrac{\vdots}{\varphi\Rightarrow\neg\bigwedge\Delta}}{\varphi,\neg\neg\bigwedge\Delta\Rightarrow}\,\mathrm{R}\neg 1}{\varphi,\Delta\Rightarrow}\,\text{Cut}$$

Ad 3. The following derivation suffices:

$$\dfrac{\dfrac{\dfrac{}{\psi_1, \psi_2 \Rightarrow \psi_1}\;\text{Ax} \qquad \dfrac{\dfrac{\vdots}{\varphi \Rightarrow \neg\psi_1}}{\varphi, \psi_1 \Rightarrow}\;\text{Item 2}}{\dfrac{\varphi, \psi_1, \psi_2 \Rightarrow}{\varphi, (\psi_1 \wedge \psi_2) \Rightarrow}\;\text{R}\wedge 2}\;\text{Cut}}{\varphi \Rightarrow \neg(\psi_1 \wedge \psi_2)}\;\text{R}\neg 2$$

Ad 4. The following derivation suffices:

$$\dfrac{\dfrac{\dfrac{\vdots}{\varphi_1 \Rightarrow \neg \bigwedge \Delta_1}}{\varphi_1 \Rightarrow \neg \bigwedge \Delta_1 \cup \Delta_2}\;\text{Item 3} \qquad \dfrac{\dfrac{\vdots}{\varphi_2 \Rightarrow \neg \bigwedge \Delta_2}}{\varphi_2 \Rightarrow \neg \bigwedge \Delta_1 \cup \Delta_2}\;\text{Item 3}}{\varphi_1 \vee \varphi_2 \Rightarrow \neg \bigwedge \Delta_1 \cup \Delta_2}\;\text{R}\vee 2$$

QED

We now prove some lemmas concerning $\vdash_{\mathcal{S}}$. Unless stated otherwise, capital Greek letters $\Delta, \Theta, \ldots$ refer to finite sets in $\mathcal{L}^{io}$. Recall that $\Delta^{\downarrow} := \{\varphi \mid \varphi^x \in \Delta \text{ for some } x \in \{o, f, c\}\}$ denotes the set of formulae Δ stripped from its labels. We write $\varphi^{\downarrow}$ denote the formula φ^x stripped from its label x. Lemma 6.1 expresses that norms, e.g., (φ, ψ) cannot occur on the rhs of a sequent (with the exception of $(\top, \top)$) and that negated norms, e.g., $\neg(\varphi, \psi)$, cannot occur on the lhs of a sequent. Lemma 6.2 states that the tautological $(\top, \top)$ only occurs on the rhs of a sequent whenever the lhs is empty. More importantly, Lemma 6.3 expresses the height-preserving invertibility of the **R-N** rule, i.e., if $\vdash^n_{\mathcal{S}} \Delta \Rightarrow \neg(\varphi, \psi)$, then $\vdash^n_{\mathcal{S}} \Delta, (\varphi, \psi) \Rightarrow$. Lemma 6.4 demonstrates that any tautological constraint ψ^c on the lhs of a sequent, i.e., $\vdash \psi^c$, can be height-preservingly eliminated from that sequent. Next, Lemma 6.5 expresses some useful facts about the conditions under which $\mathsf{DAC}_{\mathcal{S}}$ sequents stripped from their labels (cf. $\Delta^{\downarrow}$ and $\varphi^{\downarrow}$ above) correspond to derivability in the underlying base logic L. Lemma 6.6 tells us that whenever a sequent $\Delta \Rightarrow \psi^o$ concludes some obligation ψ^o, then this obligation can be height-preservingly concluded in a sequent that does not contain any constraint formula, i.e., $\Delta \setminus \mathcal{L}^c \Rightarrow \psi^o$. We end this section with the central result Lemma 6.7, which demonstrates that if a sequent $\Delta \Rightarrow$ concludes an empty rhs, this means that we can conclude from its constraint-free lhs $\Delta \setminus \mathcal{L}^c$ the negation of any obligation implied by the set of constraints on the lhs of the sequent, i.e., $\Delta \setminus \mathcal{L}^c \Rightarrow \varphi^o$ with $\varphi \vdash \neg \bigwedge(\Delta \cap \mathcal{L}^c)^{\downarrow}$. A special case is when there are no constraints on the lhs. In that case, we can derive falsum, i.e., $\Delta \Rightarrow \bot^o$. The last lemma is an invertibility lemma since it allows us to restore the obligations from which the constraints were obtained using **R-C**.

Lemma 6.1. *If* $\vdash_{\mathcal{S}} \Delta \Rightarrow \Gamma$ *then (i)* $\Delta \cap \overline{\mathcal{L}^n} = \emptyset$ *and (ii)* $\Gamma \cap (\mathcal{L}^n \setminus \{(\top, \top)\}) = \emptyset$.

Proof. The proofs are by induction on the length of the derivation of $\Delta \Rightarrow \Gamma$. Ad 1. *Base case.* No initial sequent introduces an $\overline{\mathcal{L}^n}$ formula on the lhs. *Inductive step.* None of the remaining rules introduces such a formula. Ad 2. *Base case.* No initial sequent except for **Taut** has an $\mathcal{L}^n$ formula on the rhs. *Inductive step.* No rule introduces such a formula. QED

Lemma 6.2. *If $\vdash_{\mathcal{S}} \Delta \Rightarrow (\top, \top)$ then $\Delta = \emptyset$.*

Proof. The proof is by induction on the length of the derivation of $\Delta \Rightarrow (\top, \top)$. *Base case.* The only way to introduce $\Delta \Rightarrow (\top, \top)$ is by **Taut**. *Inductive step.* **R-C** and **R-N** can be safely ignored since their consequent is not of the right form. We consider **Cut**:

$$\dfrac{\Delta_1 \Rightarrow \sigma \qquad \sigma, \Delta_2 \Rightarrow (\top, \top)}{\Delta_1, \Delta_2 \Rightarrow (\top, \top)}$$

The right premise is impossible due to the IH. The premise of **L-CT** is impossible due to the IH, and the same reasoning applies to **L-OR**. QED

Lemma 6.3. *If $\vdash^n_{\mathcal{S}} \Delta \Rightarrow \neg(\varphi, \psi)$ then $\vdash^n_{\mathcal{S}} \Delta, (\varphi, \psi) \Rightarrow$.*

Proof. The proof is by induction on the length of the derivation of $\Delta \Rightarrow \neg(\varphi, \psi)$. *Base case ($n = 1$).* No axiom produces a conclusion of the form $\neg(\varphi, \psi)$.

Inductive step ($n \mapsto n + 1$). The rule **R-C** can be safely ignored.

R-N. If $\Delta \Rightarrow \neg(\varphi, \psi)$ is derived from $\Delta, (\varphi, \psi) \Rightarrow$, then, by **R-N**, the statement holds trivially.

L-CT. Let $\Delta = \Delta_1 \cup \{\sigma^o\}$, $\Delta_1, \sigma^o \Rightarrow \neg(\varphi, \psi)$ is derived from $\Delta_1, \sigma^f \Rightarrow \neg(\varphi, \psi)$. By the IH, $\vdash^n_{\mathcal{S}} \Delta_1, \sigma^f, (\varphi, \psi) \Rightarrow$. By **L-CT**, $\vdash^{n+1}_{\mathcal{S}} \Delta_1, \sigma^o, (\varphi, \psi) \Rightarrow$.

L-OR. The case is analogous to **L-CT**.

Cut. Let $\Delta = \Delta_1 \cup \Delta_2$. Suppose

$$\dfrac{\vdash^n_{\mathcal{S}} \Delta_1 \Rightarrow \sigma \qquad \vdash^n_{\mathcal{S}} \Delta_2, \sigma \Rightarrow \neg(\varphi, \psi)}{\vdash^{n+1}_{\mathcal{S}} \Delta_1, \Delta_2 \Rightarrow \neg(\varphi, \psi)}\ \textbf{Cut}$$

By the IH, $\vdash^n_{\mathcal{S}} \Delta', \sigma, (\varphi, \psi) \Rightarrow$. By **Cut**, $\vdash^{n+1}_{\mathcal{S}} \Delta, \Delta', (\varphi, \psi) \Rightarrow$.

QED

Lemma 6.4. *If $\vdash^n_{\mathcal{S}} \Delta, \psi^c \Rightarrow \Theta$ and $\vdash \psi$, then $\vdash^n_{\mathcal{S}} \Delta \Rightarrow \Theta$.*

Proof. The proof is by induction on the length of the derivation of $\Delta, \psi^c \Rightarrow \Theta$. *Base case ($n = 1$).* Suppose $\vdash^1_{\mathcal{S}} \Delta, \psi^c \Rightarrow \Theta$. Then $\Delta, \psi^c \Rightarrow \Theta$ has been introduced by **Ax**. So, $\Delta, \Theta \subseteq \mathcal{L}^c$ and $\vdash_{\mathsf{LC}} \Delta^{\downarrow}, \psi \Rightarrow \Theta^{\downarrow}$. Since $\vdash_{\mathsf{LC}} \Rightarrow \psi$, by **Cut**, $\vdash_{\mathsf{LC}} \Delta^{\downarrow} \Rightarrow \Theta^{\downarrow}$. By **Ax**, $\vdash^1_{\mathcal{S}} \Delta \Rightarrow \Theta$.

Inductive step ($n \mapsto n+1$). Suppose $\vdash^{n+1}_{\mathcal{S}} \Delta, \psi^c \Rightarrow \Theta$. We consider the rules applied in deriving $\Delta, \psi^c \Rightarrow \Theta$.

R-C. Suppose

$$\dfrac{\vdash^n_{\mathcal{S}} \Delta \Rightarrow \varphi^o}{\vdash^{n+1}_{\mathcal{S}} \Delta, (\underbrace{\neg\varphi}_{\psi})^c \Rightarrow}$$

with $\psi = \neg\varphi$. We have:

$$\dfrac{\vdash^n_{\mathcal{S}} \Delta \Rightarrow \varphi^o \qquad \dfrac{\dfrac{\dfrac{\dfrac{\dfrac{}{\vdash_{\mathsf{LC}} \varphi \Rightarrow \varphi}}{\vdash_{\mathsf{LC}} \varphi, \neg\varphi \Rightarrow}\text{R}\neg 1}{\vdash_{\mathsf{LC}} \varphi \Rightarrow \neg\neg\varphi}\text{R}\neg 2 \quad \text{R}\neg 1\,\dfrac{\vdash_{\mathsf{LC}} \Rightarrow \overbrace{\neg\varphi}^{\psi}}{\vdash_{\mathsf{LC}} \neg\neg\varphi \Rightarrow}}{\vdash_{\mathsf{LC}} \varphi \Rightarrow}\text{Cut}}{\vdash^1_{\mathcal{S}} \varphi^o \Rightarrow}\,\mathbf{Ax}}{\vdash^{n+1}_{\mathcal{S}} \Delta \Rightarrow}\,\mathbf{Cut}$$

R-N, L-OR, and L-CT. In each case, ψ^c is not the result of the rule's application (i.e., ψ^c is not principal), and we can apply the IH to the premises.

Cut. Assume $\Delta, \psi^c \Rightarrow \Theta$ is obtained by an application of **Cut**. We have two possibilities:

$$\dfrac{\vdash^n_{\mathcal{S}} \Delta_1, \psi^c \Rightarrow \sigma \qquad \vdash^n_{\mathcal{S}} \Delta_2, \sigma \Rightarrow \Theta}{\vdash^{n+1}_{\mathcal{S}} \Delta, \psi^c \Rightarrow \Theta}\,\mathbf{Cut}$$

respectively,

$$\dfrac{\vdash_{\mathcal{S}} \Delta_1 \Rightarrow \sigma \qquad \vdash_{\mathcal{S}} \Delta_2, \psi^c, \sigma \Rightarrow \Theta}{\vdash^{n+1}_{\mathcal{S}} \Delta, \psi^c \Rightarrow \Theta}\,\mathbf{Cut}$$

By the induction hypothesis $\vdash^n_{\mathcal{S}} \Delta_1 \Rightarrow \sigma$, respectively $\vdash^n_{\mathcal{S}} \Delta_2, \sigma \Rightarrow \Theta$, and so:

$$\dfrac{\vdash^n_{\mathcal{S}} \Delta_1 \Rightarrow \sigma \qquad \vdash^n_{\mathcal{S}} \Delta_2, \sigma \Rightarrow \Theta}{\vdash^{n+1}_{\mathcal{S}} \Delta \Rightarrow \Theta}\,\mathbf{Cut}$$

QED

Lemma 6.5. *Let* $\vdash_{\mathcal{S}} \Delta \Rightarrow \gamma$. *Then,*

1. *$\Delta \subseteq \mathcal{L}^x$ and $\Delta^{\downarrow} \vdash \gamma^{\downarrow}$, if $\gamma \in \mathcal{L}^x$ with $x \in \{c, f\}$.*

2. *$\Delta^{\downarrow} \vdash \gamma^{\downarrow}$, if $\Delta \subseteq \mathcal{L}^f \cup \mathcal{L}^o \cup \{(\top, \top)\}$ and $\gamma \in \mathcal{L}^f \cup \mathcal{L}^o$.*

3. *$\vdash \gamma^{\downarrow}$, if $\Delta \subseteq \mathcal{L}^f \cup \{(\top, \top)\}$, $\gamma \in \mathcal{L}^o$, and* ***TP*** *$\notin \mathcal{S}$.*

Proof. The proofs are by induction on the length of the derivation of $\Delta \Rightarrow \gamma$. *Base case ($n = 1$).* Ad 1. In this case, $\vdash_{\mathsf{LC}} \Delta^{\downarrow} \Rightarrow \gamma^{\downarrow}$ and, by **Ax**, $\vdash^1_{\mathcal{S}} \Delta \Rightarrow \gamma$. By the adequacy of LC, $\Delta^{\downarrow} \vdash \gamma^{\downarrow}$. Ad 2. We only have the possible cases where $\Delta \Rightarrow \gamma$ is introduced by **TP**, by **Ax**, or by **Detach**. The first case follows by the reflexivity of $\vdash$, the second by the adequacy of LC. Else, it is introduced by **Detach**, where $\Delta = \{\top^f, (\top, \top)\}$ and $\gamma = \top^o$. Trivially $\vdash \gamma^{\downarrow}$. Ad 3. When $\Delta \Rightarrow \gamma$ is introduced by **Detach**, proceed as in Item 2. The only other way $\Delta \Rightarrow \gamma$ can be introduced is by **Ax**. Since $\Delta \subseteq \mathcal{L}^f$ and $\gamma \in \mathcal{L}^o$, $\Delta = \emptyset$. $\vdash \gamma^{\downarrow}$ follows by the adequacy of LC.

Inductive step ($n \mapsto n + 1$). We can exclude the cases **R-C**, **R-N**.

Cut. Consider, where $\Delta = \Delta_1 \cup \Delta_2$,

$$\dfrac{\vdash^n_{\mathcal{S}} \Delta_1 \Rightarrow \psi \qquad \vdash^n_{\mathcal{S}} \Delta_2, \psi \Rightarrow \gamma}{\vdash^{n+1}_{\mathcal{S}} \Delta_1, \Delta_2 \Rightarrow \gamma}\ \textbf{Cut}$$

By Lemma 6.1, $\psi \in \mathcal{L}^f \cup \mathcal{L}^o \cup \mathcal{L}^c \cup \{(\top, \top)\}$.

Ad 1. Since $\gamma \in \mathcal{L}^x$, by the IH, $\Delta_2 \cup \{\psi\} \subseteq \mathcal{L}^x$ and $(\Delta_2 \cup \{\psi\})^{\downarrow} \vdash \gamma^{\downarrow}$. Since $\psi \in \mathcal{L}^x$, by the IH, $\Delta_1 \subseteq \mathcal{L}^x$ with $\Delta_1^{\downarrow} \vdash \psi^{\downarrow}$. So, $\Delta \subseteq \mathcal{L}^x$ and so $\Delta^{\downarrow} \vdash \gamma^{\downarrow}$.

Ad 2. In case $\psi \in \mathcal{L}^f$, by the IH, $(\Delta_2 \cup \{\psi\})^{\downarrow} \vdash \gamma^{\downarrow}$. By Item 1, $\Delta_1^{\downarrow} \vdash \psi^{\downarrow}$ and so $\Delta^{\downarrow} \vdash \gamma^{\downarrow}$. If $\psi \in \mathcal{L}^o$, by the IH, $\Delta_1^{\downarrow} \vdash \psi^{\downarrow}$ and $(\Delta_2 \cup \{\psi\})^{\downarrow} \vdash \gamma^{\downarrow}$. So, $\Delta^{\downarrow} \vdash \gamma^{\downarrow}$. In case $\psi \in \mathcal{L}^c$, by item 1, and since $\Delta_1 \subseteq \mathcal{L}^f \cup \mathcal{L}^o$, we have $\Delta_1 = \emptyset$. By Lemma 6.4, $\vdash^n_{\mathcal{S}} \Delta_2 \Rightarrow \gamma$. By the IH, $\Delta_2^{\downarrow} \vdash \gamma^{\downarrow}$ and $\Delta^{\downarrow} = \Delta_2^{\downarrow}$. If $\psi = (\top, \top)$, by Lemma 6.2 we know $\Delta_1 = \emptyset$ and so $\Delta_2 = \Delta$. By the IH, $\Delta_2^{\downarrow} \vdash \gamma^{\downarrow}$ and so $\Delta^{\downarrow} \vdash \gamma^{\downarrow}$.

Ad 3. If $\psi \in \mathcal{L}^f \cup \{(\top, \top)\}$, by the IH, $\vdash \gamma^{\downarrow}$. Else, $\psi \in \mathcal{L}^c$. By item 1, and since $\Delta_1 \subseteq \mathcal{L}^f \cup \mathcal{L}^o \cup \{(\top, \top)\}$, we know $\Delta_1 = \emptyset$ and so $\vdash \psi^{\downarrow}$. By Lemma 6.4, $\vdash^n_{\mathcal{S}} \Delta_2^{\downarrow} \Rightarrow \gamma^{\downarrow}$. By the IH, $\vdash \gamma^{\downarrow}$.

L-OR. Suppose, where $\Delta = \Delta' \cup \{(\varphi \vee \psi)^f\}$ and $\Delta_1 \cap \mathcal{L}^n \neq \emptyset \neq \Delta_2 \cap \mathcal{L}^n$,

$$\dfrac{\vdash^n_{\mathcal{S}} \Delta_1, \varphi^f \Rightarrow \gamma \qquad \vdash^n_{\mathcal{S}} \Delta_2, \psi^f \Rightarrow \gamma}{\vdash^{n+1}_{\mathcal{S}} \Delta', (\varphi \vee \psi)^f \Rightarrow \gamma}\ \textbf{L-OR}$$

Ad 1. If $x = c$ or **TP** $\notin \mathcal{S}$, the premises are not derivable by the IH. Else, by the IH, $\Delta_1, \Delta_2 \subseteq \mathcal{L}^f$ and $\Delta_1^{\downarrow}, \varphi \vdash \gamma^{\downarrow}$ and $\Delta_2^{\downarrow}, \psi \vdash \gamma^{\downarrow}$. By $\mathrm{R}\vee 2$, $\Delta'^{\downarrow}, \varphi \vee \psi \vdash \gamma^{\downarrow}$.

Ad 2. Since $\Delta \subseteq \mathcal{L}^o \cup \mathcal{L}^f \cup \{(\top, \top)\}$, $(\Delta_1 \cap \mathcal{L}^n) \cup (\Delta_2 \cap \mathcal{L}^n) \subseteq \{(\top, \top)\}$. By the IH, $(\Delta_1 \cup \{\varphi\})^{\downarrow} \vdash \gamma^{\downarrow}$ and $(\Delta_2 \cup \{\psi\})^{\downarrow} \vdash \gamma^{\downarrow}$. By R$\vee$2, $(\Delta' \cup \{\varphi \vee \psi\})^{\downarrow} \vdash \gamma^{\downarrow}$ and so $\Delta^{\downarrow} \vdash \gamma^{\downarrow}$.

Ad 3. Since $\Delta \subseteq \mathcal{L}^o \cup \mathcal{L}^f \cup \{(\top, \top)\}$, $(\Delta_1 \cap \mathcal{L}^n) \cup (\Delta_2 \cap \mathcal{L}^n) \subseteq \{(\top, \top)\}$. By the IH, $\vdash \gamma^{\downarrow}$.

L-CT. Suppose, where $\Delta = \Delta_1 \cup \{\varphi^o\}$ and $\Delta \cap \mathcal{L}^n \neq \emptyset$,

$$\frac{\vdash^n_{\mathcal{S}} \varphi^f, \Delta_1 \Rightarrow \gamma}{\vdash^{n+1}_{\mathcal{S}} \varphi^o, \Delta_1 \Rightarrow \gamma}\ \textbf{L-CT}$$

Ad 1. By the IH, the premise cannot be derived.

Ad 2. By the IH, $\varphi, \Delta_1^{\downarrow} \vdash \gamma^{\downarrow}$ and so $\Delta^{\downarrow} \vdash \gamma^{\downarrow}$.

Ad 3. By the IH, $\vdash \gamma^{\downarrow}$. QED

Lemma 6.6. *If* $\vdash^n_{\mathcal{S}} \Delta \Rightarrow \psi^o$ *then* $\vdash^n_{\mathcal{S}} \Delta \setminus \mathcal{L}^c \Rightarrow \psi^o$.

Proof. The proof is by induction of the length of the derivation of $\Delta \Rightarrow \psi^o$. *Base case* $(n = 1)$. No sequent that can be derived in one step fits the form $\Delta, \varphi^c \Rightarrow \psi^o$.

Inductive step $(n \mapsto n+1)$. The sequents introduced by **R-C** and **R-N** have a different form and can be ignored.

L-CT. Consider,

$$\frac{\vdash^n_{\mathcal{S}} \sigma^f, \Delta \Rightarrow \psi^o}{\vdash^{n+1}_{\mathcal{S}} \sigma^o, \Delta \Rightarrow \psi^o}\ \textbf{L-CT}$$

By the IH, $\vdash^n_{\mathcal{S}} \Delta \setminus \mathcal{L}^c, \sigma^f \Rightarrow \psi^o$. By **L-CT**, $\vdash^{n+1}_{\mathcal{S}} \sigma^o, \Delta \setminus \mathcal{L}^c \Rightarrow \psi^o$.

L-OR. Consider, where $\Delta = \Delta_1 \cup \Delta_2 \cup \{(\varphi_1 \vee \varphi_2)^f\}$,

$$\frac{\vdash^n_{\mathcal{S}} \Delta_1, \varphi_1^f \Rightarrow \psi^o \qquad \vdash^n_{\mathcal{S}} \Delta_2, \varphi_2^f \Rightarrow \psi^o}{\vdash^{n+1}_{\mathcal{S}} \Delta_1, \Delta_2, (\varphi_1 \vee \varphi_2)^f \Rightarrow \psi^o}\ \textbf{L-OR}$$

By the IH, $\vdash^n_{\mathcal{S}} \Delta_1 \setminus \mathcal{L}^c, \varphi_1^f \Rightarrow \psi^o$ and $\vdash^n_{\mathcal{S}} \Delta_2 \setminus \mathcal{L}^c, \varphi_2^f \Rightarrow \psi^o$. By **L-OR**, $\vdash^{n+1}_{\mathcal{S}}$ $(\Delta_1 \cup \Delta_2) \setminus \mathcal{L}^c, (\varphi_1 \vee \varphi_2)^f \Rightarrow \psi^o$.

Cut. Consider, where $\Delta = \Delta_1 \cup \Delta_2$,

$$\frac{\vdash^n_{\mathcal{S}} \Delta_1 \Rightarrow \sigma \qquad \vdash^n_{\mathcal{S}} \sigma, \Delta_2 \Rightarrow \psi^o}{\vdash^{n+1}_{\mathcal{S}} \Delta \Rightarrow \psi^o}\ \textbf{Cut}$$

By Lemma 6.1, $\sigma \in \mathcal{L}^f \cup \mathcal{L}^o \cup \mathcal{L}^c \cup \{(\top, \top)\}$. If $\sigma \in \mathcal{L}^f$, by Lemma 6.5.1, $\Delta_1 \subseteq \mathcal{L}^f$. By the IH, $\vdash^n_{\mathcal{S}} \Delta_2 \setminus \mathcal{L}^c \Rightarrow \psi^o$. By **Cut**, $\vdash^{n+1}_{\mathcal{S}} \Delta_1 \cup (\Delta_2 \setminus \mathcal{L}^c) \Rightarrow \psi^o$ and since $\Delta_1 \cup (\Delta_2 \setminus \mathcal{L}^c) = \Delta \setminus \mathcal{L}^c$, $\vdash^{n+1}_{\mathcal{S}} \Delta \setminus \mathcal{L}^c \Rightarrow \psi^o$. If $\sigma \in \mathcal{L}^o$, by the IH, $\vdash^n_{\mathcal{S}} \Delta_1 \setminus \mathcal{L}^c \Rightarrow \sigma$ and $\vdash^n_{\mathcal{S}} \Delta_2 \setminus \mathcal{L}^c \Rightarrow \psi^o$. By **Cut**, $\vdash^{n+1}_{\mathcal{S}} \Delta \setminus \mathcal{L}^c \Rightarrow \psi^o$. If $\sigma \in \mathcal{L}^c$, by Lemma 6.5.1, $\Delta_1 \subseteq \mathcal{L}^c$. By the IH, $\vdash^n_{\mathcal{S}} \Delta_2 \setminus \mathcal{L}^c \Rightarrow \psi^o$. Since $\Delta \setminus \mathcal{L}^c = \Delta_2 \setminus \mathcal{L}^c$, $\vdash^n_{\mathcal{S}} \Delta \setminus \mathcal{L}^c \Rightarrow \psi^o$.

QED

Lemma 6.7. *If $\vdash_{\mathcal{S}} \Delta \Rightarrow$ then $\vdash_{\mathcal{S}} \Delta \setminus \mathcal{L}^c \Rightarrow \varphi^o$ such that $\varphi \vdash \neg \bigwedge(\Delta \cap \mathcal{L}^c)^{\downarrow}$ (where $\neg \bigwedge \emptyset := \bot$).*

Proof. The proof is by induction on the length of the derivation of $\Delta \Rightarrow$. *Base case ($n = 1$).* If $\vdash^1_{\mathcal{S}} \Delta \Rightarrow$ then $\Delta \Rightarrow$ is introduced by **Ax**. Suppose $\Delta \subseteq \mathcal{L}^c$. So $\vdash_{\mathsf{LC}} \Delta \Rightarrow$. By R$\wedge$1, $\vdash_{\mathsf{LC}} \bigwedge \Delta \Rightarrow$ and by R$\neg$1, $\vdash_{\mathsf{LC}} \Rightarrow \neg \bigwedge \Delta$. By **Ax**, $\vdash^1_{\mathcal{S}} \Rightarrow (\neg \bigwedge \Delta)^o$. By the reflexivity of $\vdash$, $\vdash_{\mathsf{LC}} \neg \bigwedge \Delta^{\downarrow} \Rightarrow \neg \bigwedge \Delta^{\downarrow}$. Suppose now that $\Delta \subseteq \mathcal{L}^x$ for $x \in \{f, o\}$. Hence, $\vdash_{\mathsf{LC}} \bigwedge \Delta \Rightarrow \bot$, so by **Ax** we know $\vdash^1_{\mathcal{S}} \Delta^o \Rightarrow \bot^o$ and $\bigwedge \Delta \vdash \bot$.

Inductive step ($n \mapsto n + 1$). Suppose $\vdash^{n+1}_{\mathcal{S}} \Delta \Rightarrow$.

R-C Suppose, where $\Delta = \Delta_1 \cup \{(\neg\varphi)^c\}$,

$$\dfrac{\vdash^n_{\mathcal{S}} \Delta_1 \Rightarrow \varphi^o}{\vdash^{n+1}_{\mathcal{S}} \Delta_1, (\neg\varphi)^c \Rightarrow}\ \textbf{R-C}$$

By Lemma 6.6, $\vdash^n_{\mathcal{S}} \Delta_1 \setminus \mathcal{L}^c \Rightarrow \varphi^o$ and so $\vdash^{n+1}_{\mathcal{S}} \Delta_1 \setminus \mathcal{L}^c \Rightarrow \varphi^o$. Since $\varphi \vdash \neg\neg\varphi$, $\varphi \vdash \neg \bigwedge(\Delta \cap \mathcal{L}^c)^{\downarrow}$.

R-N does not apply.

L-CT Suppose, where $\Delta = \Delta_1 \cup \{\varphi^o\}$,

$$\dfrac{\vdash^n_{\mathcal{S}} \varphi^f, \Delta_1 \Rightarrow}{\vdash^{n+1}_{\mathcal{S}} \varphi^o, \Delta_1 \Rightarrow}\ \textbf{L-CT}$$

By the IH, $\vdash_{\mathcal{S}} \Delta' \setminus \mathcal{L}^c \cup \{\varphi^f\} \Rightarrow \sigma^o$ such that $\sigma \vdash \neg \wedge (\Delta' \cap \mathcal{L}^c)^{\downarrow}$. Then:

$$\dfrac{\varphi^f, \Delta_1 \setminus \mathcal{L}^c \Rightarrow \sigma^o}{\varphi^o, \Delta_1 \setminus \mathcal{L}^c \Rightarrow \sigma^o}\ \textbf{L-CT}$$

L-OR Suppose, where $\Delta = \Delta_1 \cup \Delta_2 \cup \{(\varphi_1 \vee \varphi_2)^f\}$,

$$\dfrac{\vdash^n_{\mathcal{S}} \Delta_1, \varphi_1^f \Rightarrow \qquad \vdash^n_{\mathcal{S}} \Delta_2, \varphi_2^f \Rightarrow}{\vdash^{n+1}_{\mathcal{S}} \Delta_1, \Delta_2, (\varphi_1 \vee \varphi_2)^f \Rightarrow}\ \textbf{L-OR}$$

By the IH, there are $\sigma_1, \sigma_2 \in \mathcal{L}^p$ for which $\vdash_{\mathcal{S}} \Delta_1 \backslash \mathcal{L}^c, \varphi_1^f \Rightarrow \sigma_1^o, \vdash_{\mathcal{S}} \Delta_2 \backslash \mathcal{L}^c, \varphi_2^f \Rightarrow \sigma_2'$, $\sigma_1 \vdash \neg \bigwedge(\Delta_1 \cap \mathcal{L}^c)^{\downarrow}$ and $\sigma_2 \vdash \neg \bigwedge(\Delta_2 \cap \mathcal{L}^c)^{\downarrow}$. Then,

$$\dfrac{\dfrac{\Delta_1 \setminus \mathcal{L}^c, \varphi_1^f \Rightarrow \sigma_1^o \qquad \dfrac{\dfrac{\vdash_{\mathsf{LC}} \sigma_1 \Rightarrow \sigma_1}{\vdash_{\mathsf{LC}} \sigma_1 \Rightarrow \sigma_1 \vee \sigma_2}\,\mathrm{R}{\vee}1}{\sigma_1^o \Rightarrow (\sigma_1 \vee \sigma_2)^o}\,\mathbf{Ax}}{\Delta_1 \setminus \mathcal{L}^c, \varphi_1^f \Rightarrow (\sigma_1 \vee \sigma_2)^o}\,\mathbf{Cut} \qquad \dfrac{\vdots}{\Delta_2 \setminus \mathcal{L}^c, \varphi_2^f \Rightarrow (\sigma_1 \vee \sigma_2)^o}}{\Delta_1 \setminus \mathcal{L}^c, \Delta_2 \setminus \mathcal{L}^c, (\varphi_1 \vee \varphi_2)^f \Rightarrow (\sigma_1 \vee \sigma_2)^o}\,\mathbf{L\text{-}OR}$$

Since $\sigma_1 \vee \sigma_2 \vdash \neg \bigwedge(\Delta \cap \mathcal{L}^c)^{\downarrow}$ this completes our case.

Cut Suppose, where $\Delta = \Delta_1 \cup \Delta_2$,

$$\dfrac{\vdash_{\mathcal{S}}^{n} \Delta_1 \Rightarrow \sigma \qquad \vdash_{\mathcal{S}}^{n} \sigma, \Delta_2 \Rightarrow}{\vdash_{\mathcal{S}}^{n+1} \Delta_1, \Delta_2 \Rightarrow}\,\mathbf{Cut}$$

By the IH, $\vdash_{\mathcal{S}} (\Delta_2 \cup \{\sigma\}) \setminus \mathcal{L}^c \Rightarrow \varphi^o$ such that $\varphi \vdash \neg \bigwedge((\Delta_2 \cup \{\sigma\}) \cap \mathcal{L}^c)$. If $\sigma \in \mathcal{L}^c$, by Lemma 6.5.1, $\Delta_1 \subseteq \mathcal{L}^c$ and so $\Delta \setminus \mathcal{L}^c = \Delta_2 \setminus \mathcal{L}^c$. Since $\Delta_1^{\downarrow} \vdash \sigma^{\downarrow}$, $\varphi \vdash \neg \bigwedge(\Delta_1 \cup \Delta_2) \cap \mathcal{L}^c$. If $\sigma \in \mathcal{L}^f$ then by Lemma 6.5.1 $\Delta_1 \subseteq \mathcal{L}^f$ and so $\vdash_{\mathcal{S}}^{n} \Delta_1 \setminus \mathcal{L}^c \Rightarrow \sigma$. If $\sigma \in \mathcal{L}^o$ then by Lemma 6.6 $\vdash_{\mathcal{S}}^{n} \Delta_1 \setminus \mathcal{L}^c \Rightarrow \sigma$. In both cases, by IH, $\sigma, \Delta_2 \setminus \mathcal{L}^c \Rightarrow \varphi^o$ such that $\varphi \vdash \neg \bigwedge(\Delta_2 \cap \mathcal{L}^c)^{\downarrow}$ and we apply

$$\dfrac{\vdash_{\mathcal{S}}^{n} \Delta_1 \setminus \mathcal{L}^c \Rightarrow \sigma \qquad \vdash_{\mathcal{S}}^{n} \sigma, \Delta_2 \setminus \mathcal{L}^c \Rightarrow \varphi^o}{\vdash_{\mathcal{S}}^{n+1} \Delta_1 \setminus \mathcal{L}^c, \Delta_2 \setminus \mathcal{L}^c \Rightarrow \varphi^o}\,\mathbf{Cut}$$

QED

6.4.2 Soundness and Completeness: DAC and deriv

This section provides the proof of soundness and completeness between I/O proof systems deriv and our DAC systems (Theorem 6.1). The assumed correspondence between the rules of deriv and those of DAC are presented in Table 6.1. First, we prove a useful result concerning deriv: Lemma 6.8 can be interpreted as proving a transitivity property of deriv when merging two sets of norms.

Lemma 6.8. *Let $\Theta_1, \Theta_2 \subseteq \mathcal{L}^n$. If $(\gamma_1, \varphi) \in \mathsf{deriv}_{\mathcal{R},\mathsf{L}}(\Theta_1)$ and $(\gamma_2, \psi) \in \mathsf{deriv}_{\mathcal{R},\mathsf{L}}(\Theta_2)$ such that $(\top, \varphi) \in \Theta_2$ and $\gamma_1 \vdash \gamma_2$, then $(\gamma_1, \psi) \in \mathsf{deriv}_{\mathcal{R},\mathsf{L}}(\Theta^*)$ with $\Theta^* = \Theta_1 \cup (\Theta_2 \setminus \{(\top, \varphi)\})$.*

Proof. The proof is by induction on the length of the derivation of $(\gamma_2, \psi) \in \mathsf{deriv}_{\mathcal{R},\mathsf{L}}(\Theta_2)$.

Base case. In this case, $(\top, \varphi) = (\gamma_2, \psi)$ and $\{(\top, \varphi)\} = \Theta_2$. Since $(\gamma_1, \varphi) \in \mathsf{deriv}_{\mathcal{R},\mathsf{L}}(\Theta_1)$ we have $(\gamma_1, \varphi) \in \mathsf{deriv}_{\mathcal{R},\mathsf{L}}(\Theta^*)$ since $\Theta_1 = \Theta^*$ (nb. $\gamma_1 \vdash \top$).

Inductive step. We consider the rules through which (ψ_2, ψ) can be derived.

Rules of $\mathsf{deriv}_{\mathcal{R},\mathsf{L}}$	Rules of $\mathsf{DAC}_{\mathcal{S}}$
{WO, AND, SI}	$\{\mathbf{Ax}, \mathbf{Detach}, \mathbf{R\text{-}C}, \mathbf{R\text{-}N}, \mathbf{Cut}\}$
T	**Taut**
ID	**TP**
CT	**L-CT**
OR	**L-OR**

Table 6.1: Correspondence between $\mathsf{deriv}_{\mathcal{R},\mathsf{L}}$ rules and $\mathsf{DAC}_{\mathcal{S}}$ rules. E.g., $\{\text{ID}, \text{OR}\} \subseteq \mathcal{R}$ iff $\{\mathbf{TP}, \mathbf{L\text{-}OR}\} \subseteq \mathcal{S}$. The first row represents the minimal set of rules both systems must satisfy. We use boldface to distinguish DAC rules from deriv rules.

SI. Let $(\gamma_2, \psi) \in \mathsf{deriv}_{\mathcal{R},\mathsf{L}}(\Theta_2)$ through an application of SI to (γ_3, ψ) and $\gamma_2 \vdash \gamma_3$. We know $\gamma_1 \vdash \gamma_3$ and, by the IH, we obtain $(\gamma_1, \psi) \in \mathsf{deriv}_{\mathcal{R},\mathsf{L}}(\Theta^*)$.

WO. Let $(\gamma_2, \psi) \in \mathsf{deriv}_{\mathcal{R},\mathsf{L}}(\Theta_2)$ through an application of WO to (γ_2, ψ_2) and $\psi_2 \vdash \psi$. We apply the IH to the premise and obtain $(\gamma_1, \psi_2) \in \mathsf{deriv}_{\mathcal{R},\mathsf{L}}(\Theta^*)$, then we apply WO and obtain $(\gamma_1, \psi) \in \mathsf{deriv}_{\mathcal{R},\mathsf{L}}(\Theta^*)$.

AND. Let $(\gamma_2, \psi) \in \mathsf{deriv}_{\mathcal{R},\mathsf{L}}(\Theta_2)$ through an application of AND to $(\gamma_2, \psi_1) \in \mathsf{deriv}_{\mathcal{R},\mathsf{L}}(\Theta_2^1)$ and $(\gamma_2, \psi_2) \in \mathsf{deriv}_{\mathcal{R},\mathsf{L}}(\Theta_2^2)$, where $\psi = \psi_1 \wedge \psi_2$ and $\Theta_2 = \Theta_2^1 \cup \Theta_2^2$. By the IH, $(\gamma_1, \psi_1) \in \mathsf{deriv}_{\mathcal{R},\mathsf{L}}(\Theta^{1,*})$ and $(\gamma_1, \psi_1) \in \mathsf{deriv}_{\mathcal{R},\mathsf{L}}(\Theta^{2,*})$, where $\Theta^{j,*} = \Theta_1 \cup (\Theta_2^j \setminus \{(\top, \varphi)\})$ (for $j \in \{1, 2\}$). By AND, $(\gamma_1, \psi) \in \mathsf{deriv}_{\mathcal{R},\mathsf{L}}(\Theta^*)$.

OR. Let $(\gamma_2, \psi) \in \mathsf{deriv}_{\mathcal{R},\mathsf{L}}(\Theta_2)$ with $\gamma_2 = \gamma_3 \vee \gamma_4$ through an application of OR to $(\gamma_3, \psi) \in \mathsf{deriv}_{\mathcal{R},\mathsf{L}}(\Theta_2^1)$ and $(\gamma_4, \psi) \in \mathsf{deriv}_{\mathcal{R},\mathsf{L}}(\Theta_2^2)$, where $\Theta_2 = \Theta_2^1 \cup \Theta_2^2$. By R$\vee$2, $\gamma_3 \vdash \gamma_2$ and $\gamma_4 \vdash \gamma_2$. By SI and R$\wedge$1, $(\gamma_1 \wedge \gamma_3, \varphi), (\gamma_1 \wedge \gamma_4, \varphi) \in \mathsf{deriv}_{\mathcal{R},\mathsf{L}}(\Theta_1)$. By the IH, $(\gamma_1 \wedge \gamma_3, \psi) \in \mathsf{deriv}_{\mathcal{R},\mathsf{L}}(\Theta^{1,*})$ and $(\gamma_1 \wedge \gamma_4, \psi) \in \mathsf{deriv}_{\mathcal{R},\mathsf{L}}(\Theta^{2,*})$, where $\Theta^{j,*} = \Theta_1 \cup (\Theta_2^j \setminus \{(\top, \varphi)\})$ (for $j \in \{1, 2\}$). By OR, $((\gamma_1 \wedge \gamma_3) \vee (\gamma_1 \wedge \gamma_4), \psi) \in \mathsf{deriv}_{\mathcal{R},\mathsf{L}}(\Theta^*)$. By D2, $\gamma_1 \wedge \gamma_2 \vdash (\gamma_1 \wedge \gamma_3) \vee (\gamma_1 \wedge \gamma_4)$ and by R$\wedge$1, $\gamma_1 \vdash \gamma_1 \wedge \gamma_2$. Since $\vdash$ is transitive, $\gamma_1 \vdash (\gamma_1 \wedge \gamma_3) \vee (\gamma_1 \wedge \gamma_4)$. By SI, $(\gamma_1, \psi) \in \mathsf{deriv}_{\mathcal{R},\mathsf{L}}(\Theta^*)$.

CT. Let $(\gamma_2, \psi) \in \mathsf{deriv}_{\mathcal{R},\mathsf{L}}(\Theta_2)$ through an application of CT to $(\gamma_2, \psi_2) \in \mathsf{deriv}_{\mathcal{R},\mathsf{L}}(\Theta_2^1)$ and $(\gamma_2 \wedge \psi_2, \psi) \in \mathsf{deriv}_{\mathcal{R},\mathsf{L}}(\Theta_2^2)$, where $\Theta_2 = \Theta_2^1 \cup \Theta_2^2$. For $j \in \{1, 2\}$, let $\Theta^{*,j} = \Theta_1 \cup (\Theta_2^j \setminus \{(\top, \psi)\})$. By IH, $(\gamma_1, \psi_2) \in \mathsf{deriv}_{\mathcal{R},\mathsf{L}}(\Theta_1^{*,1})$. By SI, $(\gamma_1 \wedge \psi_2, \varphi) \in \mathsf{deriv}_{\mathcal{R},\mathsf{L}}(\Theta_1)$. By the IH and since $\gamma_1 \wedge \psi_2 \vdash \gamma_2 \wedge \psi_2$, $(\gamma_1 \wedge \psi_2, \psi) \in \mathsf{deriv}_{\mathcal{R},\mathsf{L}}(\Theta^{*,2})$. By CT, $(\gamma_1, \psi) \in \mathsf{deriv}_{\mathcal{R},\mathsf{L}}(\Theta^*)$.

QED

The following lemma shows the admissibility of $(\top, \top)$ whenever $T \in \mathcal{R}$ of $\mathsf{deriv}_{\mathcal{R},\mathsf{L}}$.

Lemma 6.9. *Let $\Theta \subseteq \mathcal{L}^n$ and $T \in \mathcal{R}$. If $(\varphi, \psi) \in \mathsf{deriv}_{\mathcal{R},\mathsf{L}}(\Theta \cup \{(\top, \top)\})$ then $(\varphi, \psi) \in \mathsf{deriv}_{\mathcal{R},\mathsf{L}}(\Theta)$.*

Proof. We show this by induction on the length of the proof of (φ, ψ).

Base case. In this case $\{(\varphi, \psi)\} = \Theta \cup \{(\top, \top)\}$. Clearly, by T, $(\varphi, \psi) \in \mathsf{deriv}_{\mathcal{R},\mathsf{L}}(\emptyset)$.

Inductive step. We consider the rules through which (φ, ψ) can be derived.

WO. Suppose (φ, ψ) has been derived from $(\varphi, \psi') \in \mathsf{deriv}_{\mathcal{R},\mathsf{L}}(\Theta \cup \{(\top, \top)\})$ by WO. By the IH, $(\varphi, \psi') \in \mathsf{deriv}_{\mathcal{R},\mathsf{L}}(\Theta)$ and so $(\varphi, \psi) \in \mathsf{deriv}_{\mathcal{R},\mathsf{L}}(\Theta)$.

SI. The case for SI is analogous to that of WO.

AND. Suppose (φ, ψ) has been derived from $(\varphi, \psi_1) \in \mathsf{deriv}_{\mathcal{R},\mathsf{L}}(\Theta_1)$ and $(\varphi, \psi_2) \in \mathsf{deriv}_{\mathcal{R},\mathsf{L}}(\Theta_2)$ where $\Theta \cup \{(\top, \top)\} = \Theta_1 \cup \Theta_2$ and $\psi = \psi_1 \wedge \psi_2$. By the IH, $(\varphi, \psi_1) \in \mathsf{deriv}_{\mathcal{R},\mathsf{L}}(\Theta_1 \setminus \{(\top, \top)\})$ and $(\varphi, \psi_2) \in \mathsf{deriv}_{\mathcal{R},\mathsf{L}}(\Theta_2 \setminus \{(\top, \top)\})$. So, $(\varphi, \psi) \in \mathsf{deriv}_{\mathcal{R},\mathsf{L}}(\Theta)$.

OR and CT. The cases for OR and CT are analogous to that of AND.

QED

We prove the two adequacy directions separately. For the Left-to-Right direction, we prove a slightly stronger result:

Lemma 6.10. *Let $\Omega \subseteq \mathcal{L}^p$, $\Theta \subseteq \mathcal{L}^n$, and $\Omega^{\uparrow} = \{(\top, \varphi) \mid \varphi \in \Omega\}$. If $\vdash_{\mathcal{S}} \Delta^f, \Theta, \Omega^o \Rightarrow \varphi^o$, then $\varphi \in \mathsf{deriv}_{\mathcal{R},\mathsf{L}}(\Delta, \Theta \cup \Omega^{\uparrow})$.*

Proof. We show this by induction on the length of the proof of $\Delta^f, \Theta, \Omega^o \Rightarrow \varphi^o$.

Base case. We consider the axioms.

Ax. $\vdash_{\mathcal{S}} \Omega^o \Rightarrow \varphi^o$, $\Delta = \Theta = \emptyset$ and $\Omega \vdash \varphi$. By AND, $\bigwedge \Omega \in \mathsf{deriv}_{\mathcal{R},\mathsf{L}}(\emptyset, \Omega^{\uparrow})$ and by WO, $\varphi \in \mathsf{deriv}_{\mathcal{R},\mathsf{L}}(\emptyset, \Omega^{\uparrow})$.

Detach. $\vdash_{\mathcal{S}} \psi^f, (\psi, \varphi) \Rightarrow \varphi^o$. Clearly, $\varphi \in \mathsf{deriv}_{\mathcal{R},\mathsf{L}}(\{\psi\}, \{(\psi, \varphi)\})$.

Taut. Nothing to show.

TP. $\vdash_{\mathcal{S}} \varphi^f \Rightarrow \varphi^o$. In this case, $(\varphi, \varphi) \in \mathsf{deriv}_{\mathcal{R},\mathsf{L}}(\emptyset)$ by ID, and hence $\varphi \in \mathsf{deriv}_{\mathcal{R},\mathsf{L}}(\{\varphi\}, \emptyset)$.

Inductive step. The rules **R-C** and **R-N** can be ignored.

Cut. $\Theta, \Delta^f, \Omega^o \Rightarrow \varphi^o$ is derived by **Cut** from $\Theta_1, \Delta_1^f, \Omega_1^o \Rightarrow \sigma$ and $\Theta_2, \Delta_2^f, \Omega_2^o, \sigma \Rightarrow \varphi^o$, where $\Theta = \Theta_1 \cup \Theta_2$, $\Delta = \Delta_1 \cup \Delta_2$, and $\Omega = \Omega_1 \cup \Omega_2$. By Lemma 6.1, we need to consider the cases (1) $\sigma = \psi^f \in \mathcal{L}^f$, (2) $\sigma = \psi^o \in \mathcal{L}^o$, (3) $\sigma = \psi^c \in \mathcal{L}^c$, and (4) $\sigma = (\top, \top)$.

1. By Lemma 6.5.1, $\Theta_1 \cup \Omega_1 = \emptyset$ and (†) $\vdash \Delta_1^{\downarrow} \Rightarrow \sigma^{\downarrow}$. By the IH, $\varphi \in \mathsf{deriv}_{\mathcal{R},\mathsf{L}}(\Delta_2 \cup \{\sigma\}, \Theta_2 \cup \Omega_2^{\uparrow})$ and so $(\bigwedge \Delta_2 \cup \{\sigma\}, \varphi) \in \mathsf{deriv}_{\mathcal{R},\mathsf{L}}(\Theta \cup \Omega^{\uparrow})$. By SI and (†), $(\bigwedge \Delta, \varphi) \in \mathsf{deriv}_{\mathcal{R},\mathsf{L}}(\Theta \cup \Omega^{\uparrow})$ and so $\varphi \in \mathsf{deriv}_{\mathcal{R},\mathsf{L}}(\Delta, \Theta \cup \Omega^{\uparrow})$.
2. By the IH, $\psi \in \mathsf{deriv}_{\mathcal{R},\mathsf{L}}(\Delta_1, \Theta_1 \cup \Omega_1^{\uparrow})$ and $\varphi \in \mathsf{deriv}_{\mathcal{R},\mathsf{L}}(\Delta_2, \Theta_2 \cup (\Omega_2 \cup \{\psi\})^{\uparrow})$. So, $(\bigwedge \Delta_1, \psi) \in \mathsf{deriv}_{\mathcal{R},\mathsf{L}}(\Theta_1 \cup \Omega_1^{\uparrow})$ and $(\bigwedge \Delta_2, \varphi) \in \mathsf{deriv}_{\mathcal{R},\mathsf{L}}(\Theta_2 \cup (\Omega_2 \cup \{\psi\})^{\uparrow})$. By SI, $(\bigwedge \Delta, \psi) \in \mathsf{deriv}_{\mathcal{R},\mathsf{L}}(\Theta_1 \cup \Omega_1^{\uparrow})$ and $(\bigwedge \Delta, \varphi) \in \mathsf{deriv}_{\mathcal{R},\mathsf{L}}(\Theta_2 \cup (\Omega_2 \cup \{\psi\})^{\uparrow})$. By Lemma 6.8 and since $\bigwedge \Delta \vdash \bigwedge \Delta$, $(\bigwedge \Delta, \varphi) \in \mathsf{deriv}_{\mathcal{R},\mathsf{L}}(\Theta \cup \Omega^{\uparrow})$ and so $\varphi \in \mathsf{deriv}_{\mathcal{R},\mathsf{L}}(\Delta, \Theta \cup \Omega^{\uparrow})$.
3. By Lemma 6.5.1, $\vdash \psi$ and $\Theta_1 \cup \Delta_1 \cup \Omega_1 = \emptyset$. By Lemma 6.4, $\Theta_2, \Delta_2^f, \Omega_2^o \Rightarrow \varphi^o$ is derivable in a proof of the same length as the one of $\Theta_2, \Delta_2^f, \Omega_2^o, \sigma \Rightarrow \varphi^o$. By the IH, we have $\varphi \in \mathsf{deriv}_{\mathcal{R},\mathsf{L}}(\Theta \cup \Omega^{\uparrow}, \Delta)$.
4. By Lemma 6.2, $\Theta_1 \cup \Delta_1 \cup \Omega_1 = \emptyset$ and so **Taut** $\in \mathcal{S}$. Consequently, by Table 6.1, $T \in \mathcal{R}$ too. By the IH, $\varphi \in \mathsf{deriv}_{\mathcal{R},\mathsf{L}}(\Delta, \Theta \cup \{(\top, \top)\} \cup \Omega^{\uparrow})$. By Lemma 6.9, $\varphi \in \mathsf{deriv}_{\mathcal{R},\mathsf{L}}(\Delta, \Theta \cup \Omega^{\uparrow})$.

L-CT. $\Delta^f, \Theta, \Omega^o, \psi^o \Rightarrow \varphi^o$ is derived from $\Delta^f, \Theta, \Omega^o, \psi^f \Rightarrow \varphi^o$. By the IH, $\varphi \in \mathsf{deriv}_{\mathcal{R},\mathsf{L}}(\Delta \cup \{\psi\}, \Theta \cup \Omega^{\uparrow})$ and so $(\bigwedge \Delta \cup \{\psi\}, \varphi) \in \mathsf{deriv}_{\mathcal{R},\mathsf{L}}(\Theta \cup \Omega^{\uparrow})$. By SI, from $(\top, \psi)$ follows $(\bigwedge \Delta, \psi)$. Together with $(\bigwedge \Delta \cup \{\psi\}, \varphi)$ and CT, $(\bigwedge \Delta, \varphi) \in \mathsf{deriv}_{\mathcal{R},\mathsf{L}}(\Theta \cup \Omega^{\uparrow} \cup \{(\top, \psi)\})$ and so $\varphi \in \mathsf{deriv}_{\mathcal{R},\mathsf{L}}(\Delta, \Theta \cup (\Omega \cup \{\psi\})^{\uparrow})$.

L-OR. $\Delta^f, \Theta, \Omega^o, (\sigma \vee \psi)^f \Rightarrow \varphi^o$ is derived from $\Delta_1^f, \Theta_1, \Omega_1^o, \sigma^f \Rightarrow \varphi^o$ and $\Delta_2^f, \Theta_2, \Omega_2^o, \psi^f \Rightarrow \varphi^o$, where $\Delta = \Delta_1 \cup \Delta_2$, $\Theta = \Theta_1 \cup \Theta_2$, and $\Omega = \Omega_1 \cup \Omega_2$. By the IH, $\varphi \in \mathsf{deriv}_{\mathcal{R},\mathsf{L}}(\Delta_1 \cup \{\sigma\}, \Theta_1 \cup \Omega_1^{\uparrow})$ and $\varphi \in \mathsf{deriv}_{\mathcal{R},\mathsf{L}}(\Delta_2 \cup \{\psi\}, \Theta_2 \cup \Omega_2^{\uparrow})$. So, by SI, $(\bigwedge \Delta \cup \{\sigma\}, \varphi) \in \mathsf{deriv}_{\mathcal{R},\mathsf{L}}(\Theta_1 \cup \Omega_1^{\uparrow})$ and $(\bigwedge \Delta \cup \{\psi\}, \varphi) \in \mathsf{deriv}_{\mathcal{R},\mathsf{L}}(\Theta_2 \cup \Omega_2^{\uparrow})$. By OR, $((\bigwedge \Delta \cup \{\sigma\}) \vee (\bigwedge \Delta \cup \{\psi\}), \varphi) \in \mathsf{deriv}_{\mathcal{R},\mathsf{L}}(\Theta \cup \Omega^{\uparrow})$. By SI and D1, $(\bigwedge \Delta \wedge (\sigma \vee \psi), \varphi) \in \mathsf{deriv}_{\mathcal{R},\mathsf{L}}(\Theta \cup \Omega^{\uparrow})$ and so $\varphi \in \mathsf{deriv}_{\mathcal{R},\mathsf{L}}(\Delta, \Theta \cup \Omega^{\uparrow})$.

QED

Lemma 6.11. *Let* $\Theta \subseteq \mathcal{L}^n$*. If* $(\varphi, \psi) \in \mathsf{deriv}_{\mathcal{R},\mathsf{L}}(\Theta)$ *then* $\vdash_{\mathcal{S}} \varphi^f, \Theta \Rightarrow \psi^o$*.*

Proof. We show this inductively over the length of the $\mathsf{deriv}_{\mathcal{R},\mathsf{L}}$-proof of (φ, ψ).

Base case. If $\{(\varphi, \psi)\} = \Theta$, then by **Detach**, $\vdash_{\mathcal{S}} \varphi^f, (\varphi, \psi) \Rightarrow \psi^o$. If $(\top, \top)$ is derived by T with $\Theta = \emptyset$, then by **Detach**, $\vdash_{\mathcal{S}} \top^f, (\top, \top) \Rightarrow \top^o$ and by **Taut**, $\vdash_{\mathcal{S}} \Rightarrow (\top, \top)$ and by **Cut**, $\vdash_{\mathcal{S}} \top^f \Rightarrow \top^o$. If (φ, φ) is derived by ID with $\Theta = \emptyset$, then by **TP**, $\varphi^f \Rightarrow \varphi^o$.

Inductive step. We consider all the rules through which (φ, ψ) can be derived.

WO. (φ, ψ) is derived from $(\varphi, \sigma) \in \mathsf{deriv}^i_{\mathcal{R},\mathsf{L}}(\Theta)$ and $\sigma \vdash \psi$ by WO. By the IH, $\vdash_{\mathcal{S}} \varphi^f, \Theta \Rightarrow \sigma^o$. Since $\sigma \vdash \psi$, $\vdash_{\mathcal{S}} \sigma^o \Rightarrow \psi^o$ by **Ax**. By **Cut**, $\vdash_{\mathcal{S}} \varphi^f, \Theta \Rightarrow \psi^o$.

SI. (φ, ψ) is derived from $(\sigma, \psi) \in \mathsf{deriv}_{\mathcal{R},\mathsf{L}}(\Theta)$ and $\varphi \vdash \sigma$ by SI. By the IH, $\vdash_{\mathcal{S}} \sigma^f, \Theta \Rightarrow \psi^o$. Since $\varphi \vdash \sigma$, by **Ax**, $\vdash_{\mathcal{S}} \varphi^f \Rightarrow \sigma^f$. By **Cut**, $\vdash_{\mathcal{S}} \varphi^f, \Theta \Rightarrow \psi^o$.

AND. (φ, ψ) is derived from $(\varphi, \psi_1) \in \mathsf{deriv}_{\mathcal{R},\mathsf{L}}(\Theta_1)$ and $(\varphi, \psi_2) \in \mathsf{deriv}_{\mathcal{R},\mathsf{L}}(\Theta_2)$, where $\Theta = \Theta_1 \cup \Theta_2$ and $\psi = \psi_1 \wedge \psi_2$. By the IH, $\vdash_{\mathcal{S}} \varphi^f, \Theta_1 \Rightarrow \psi_1^o$ and $\vdash_{\mathcal{S}} \varphi^f, \Theta_2 \Rightarrow \psi_2^o$. By R$\wedge$2, $\psi_1, \psi_2 \vdash \psi$. By **Ax**, $\vdash_{\mathcal{S}} \psi_1^o, \psi_2^o \Rightarrow \psi^o$. By two applications of **Cut**, $\vdash_{\mathcal{S}} \varphi^f, \Theta \Rightarrow \psi^o$.

CT. (φ, ψ) is derived from $(\varphi, \sigma) \in \mathsf{deriv}_{\mathcal{R},\mathsf{L}}(\Theta_1)$ and $(\varphi \wedge \sigma, \psi) \in \mathsf{deriv}_{\mathcal{R},\mathsf{L}}(\Theta_2)$ by CT, where $\Theta = \Theta_1 \cup \Theta_2$. By the IH, $\vdash_{\mathcal{S}} \varphi^f, \Theta_1 \Rightarrow \sigma^o$ and $\vdash_{\mathcal{S}} (\varphi \wedge \sigma)^f, \Theta_2 \Rightarrow \psi^o$. By R$\wedge$2, $\varphi, \sigma \vdash \varphi \wedge \sigma$. By **Ax**, $\vdash_{\mathcal{S}} \varphi^f, \sigma^f \Rightarrow (\varphi \wedge \sigma)^f$ and by **Cut**, $\varphi^f, \sigma^f, \Theta_2 \Rightarrow \psi^o$.

1. If $\Theta_2 \neq \emptyset$, by **L-CT**, $\vdash_{\mathcal{S}} \varphi^f, \sigma^o, \Theta_2 \Rightarrow \psi^o$ and by **Cut**, $\vdash_{\mathcal{S}} \varphi^f, \Theta \Rightarrow \psi^o$.
2. If $\Theta_2 = \emptyset$ (and hence $\Theta = \Theta_1$), we consider: (i) **TP** $\in \mathcal{S}$ and (ii) **TP** $\notin \mathcal{S}$. Ad (i). By Lemma 6.5.2, $\varphi, \sigma \vdash \psi$ and by **Ax**, $\vdash_{\mathcal{S}} \varphi^o, \sigma^o \Rightarrow \psi^o$. By **TP**, $\vdash_{\mathcal{S}} \varphi^f \Rightarrow \varphi^o$ and by twice **Cut**, $\vdash_{\mathcal{S}} \varphi^f, \Theta \Rightarrow \psi^o$. Ad (ii). By Lemma 6.5.3, $\vdash \psi$ and so $\sigma \vdash \psi$ by monotonicity. By **Ax**, $\vdash_{\mathcal{S}} \sigma^o \Rightarrow \psi^o$. By **Cut**, $\vdash_{\mathcal{S}} \varphi^f, \Theta \Rightarrow \psi^o$.

OR. (φ, ψ) is derived from $(\varphi_1, \psi) \in \mathsf{deriv}_{\mathcal{R},\mathsf{L}}(\Theta_1)$ and $(\varphi_2, \psi) \in \mathsf{deriv}_{\mathcal{R},\mathsf{L}}(\Theta_2)$, where $\varphi = \varphi_1 \vee \varphi_2$ and $\Theta = \Theta_1 \cup \Theta_2$. By the IH, $\vdash_{\mathcal{S}} \varphi_1, \Theta_1 \Rightarrow \psi^o$ and $\vdash_{\mathcal{S}} \varphi_2, \Theta_2 \Rightarrow \psi^o$. If $\Theta_1 \neq \emptyset \neq \Theta_2$, by **L-OR**, $\vdash_{\mathcal{S}} \varphi^f, \Theta \Rightarrow \psi^o$. Else, $\Theta_1 = \emptyset$ or $\Theta_2 = \emptyset$. Wlog suppose $\Theta_1 = \emptyset$. We consider: (i) **TP** $\in \mathcal{S}$ or (ii) **TP** $\notin \mathcal{S}$. Ad (i). By **L-OR**, $\vdash_{\mathcal{S}} \varphi^f, \Theta \Rightarrow \psi^o$. Ad (ii). Since $\vdash_{\mathcal{S}} \varphi_1^f \Rightarrow \psi^o$, by Lemma 6.5.3, $\vdash \psi$. By **Detach**, $\vdash_{\mathcal{S}} \top^f, (\top, \top) \Rightarrow \top^o$. Since $\vdash_{\mathcal{S}} \top^o \Rightarrow \psi^o$, $\vdash_{\mathcal{S}} \top^f, (\top, \top) \Rightarrow \psi^o$ by **Cut**. Since $\vdash_{\mathcal{S}} \varphi_1^f \Rightarrow \top^f$, by **Cut**, $\vdash_{\mathcal{S}} \varphi_1^f, (\top, \top) \Rightarrow \psi^o$. If $\Theta_2 = \emptyset$, by the same reasoning $\vdash_{\mathcal{S}} \varphi_2^f, (\top, \top) \Rightarrow \psi^o$. By **L-OR**, $\varphi^f, (\top, \top) \Rightarrow \psi^o$. Else, by **L-OR**, $\vdash_{\mathcal{S}} \varphi^f, \Theta_2, (\top, \top) \Rightarrow \psi^o$. By **Taut**, $\vdash_{\mathcal{S}} \Rightarrow (\top, \top)$ and by **Cut** $\vdash_{\mathcal{S}} \varphi^f, \Theta \Rightarrow \psi^o$ in both cases.

QED

Theorem 6.1 (Soundness and Completeness)**.** *Let $\Delta \subseteq \mathcal{L}$, $\psi \in \mathcal{L}$, and $\Theta \subseteq \mathcal{L}^n$ (where $\mathcal{L}$ is the non-labelled language from Definition 6.1). Then, $\vdash_{\mathcal{S}} \Delta^f, \Theta \Rightarrow \psi^o$ iff $\psi \in \mathsf{deriv}_{\mathcal{R},\mathsf{L}}(\Delta, \Theta)$.*

Proof. **Left-to-Right.** This is Lemma 6.10 with $\Omega = \emptyset$. **Right-to-Left.** Suppose $\psi \in \mathsf{deriv}_{\mathcal{R},\mathsf{L}}(\Delta, \Theta)$. So, $(\bigwedge \Delta, \psi) \in \mathsf{deriv}_{\mathcal{R},\mathsf{L}}(\Theta)$. By Lemma 6.11, $\vdash_{\mathcal{S}} (\bigwedge \Delta)^f, \Theta \Rightarrow \psi^o$. Since $\Delta \vdash \bigwedge \Delta$, $\vdash_{\mathcal{S}} \Delta^f \Rightarrow (\bigwedge \Delta)^f$ by **Ax**. By **Cut**, $\vdash_{\mathcal{S}} \Delta^f, \Theta \Rightarrow \psi^o$. QED

6.5 Formal Argumentation and Explanation

So far, we have shown that Deontic Argumentation Calculi are sound and complete with respect to the class of monotonic I/O proof systems deriv. However, as demonstrated in Section 6.3, these calculi generate more than arguments concluding obligations. The central DAC arguments are of two types: they either give reasons for obligations or they give reasons for why certain norms are inapplicable.[11] The latter type captures

[11] Other DAC arguments include arguments about facts and constraints.

the defeasibility of normative reasoning and defines the interaction among arguments. Namely, an argument concluding $\neg(\varphi, \psi)$, defeats all arguments making an appeal to (φ, ψ) in their reasons (see page 217). We define DAC-induced *argumentation frameworks* to model this interaction. In fact, the two types of arguments are sufficient for an argumentative characterization of *nonmonotonic* I/O logics when instantiating argumentation frameworks with DAC-arguments. This result—corresponding to Objective 3—is formally proven in Section 6.6.

An Argumentation Framework (AF) (Dung, 1995) is a tuple $\langle \mathsf{Arg}, \mathsf{Att} \rangle$ consisting of a (denumerable) set Arg of arguments $a, b, c, \ldots$, and a binary relation $\mathsf{Att} \subseteq \mathsf{Arg} \times \mathsf{Arg}$ representing defeats between these arguments.[12] In Definition 6.9 below, we define DAC-based representations of $\mathcal{AF}$s. These are $\mathcal{AF}$s instantiated with DAC-arguments.

Definition 6.9 (DAC-induced Argumentation Frameworks)**.** *Let* $\mathsf{DAC}_{\mathcal{S}}$ *be a calculus and* $\mathcal{K} = \langle \mathcal{F}, \mathcal{N}, \mathcal{C} \rangle$ *a labelled knowledge base (i.e.,* $\mathcal{F} \subseteq \mathcal{L}^f, \mathcal{N} \subseteq \mathcal{L}^n$*, and* $\mathcal{C} \subseteq \mathcal{L}^c$*). We define a* $\mathsf{DAC}_{\mathcal{S}}$-induced argumentation framework $\mathcal{AF}_{\mathcal{S}}(\mathcal{K}) = \langle \mathsf{Arg}, \mathsf{Att} \rangle$ *as follows:*

- $\Delta \Rightarrow \Gamma \in \mathsf{Arg}$ *iff* $\Delta \Rightarrow \Gamma$ *is* $\mathsf{DAC}_{\mathcal{S}}$*-derivable and* $\Delta \subseteq \mathcal{F} \cup \mathcal{N} \cup \mathcal{C}$*.*

Let $a, b \in \mathsf{Arg}$*,*

- a defeats b*, i.e.,* $(a, b) \in \mathsf{Att}$ *iff* $a = \Delta \Rightarrow \neg(\varphi, \psi)$ *and* $b = \Gamma, (\varphi, \psi) \Rightarrow \Theta$.[13]

We write $\mathsf{Arg}(\Sigma)$ *to denote the set of* $\mathsf{DAC}_{\mathcal{S}}$*-arguments* $\Delta \Rightarrow \Gamma$ *for which* $\Delta \subseteq \Sigma \subseteq \mathcal{L}^{io}$*.*

For a $\mathsf{DAC}_{\mathcal{S}}$-induced $\mathcal{AF}_{\mathcal{S}}(\mathcal{K})$ it suffices to only consider arguments relevant to the given knowledge base $\mathcal{K}$, i.e., the set of arguments $\mathsf{Arg}(\mathcal{F} \cup \mathcal{N} \cup \mathcal{C})$. We sometimes say that an argument $\Delta \Rightarrow \Gamma$ is *triggered* by the knowledge base $\mathcal{K}$ whenever $\Delta \subseteq \mathcal{F} \cup \mathcal{N} \cup \mathcal{C}$.[14]

We are interested in which arguments can be accepted given an $\mathcal{AF}$. Different semantics are available to determine arguments' acceptability (Baroni et al., 2011). These semantics give rise to extensions, i.e., collections of arguments that can be jointly defended from attacking arguments. For our purpose, *stable semantics* suffice.

[12] Various extensions of argumentation frameworks have been proposed, capturing support relations, preference-based defeats, collective defeats, and more. For our purpose, the basic defeat relation suffices. Furthermore, one may distinguish between abstract and structured argumentation (Prakken, 2018). In abstract argumentation (Dung, 1995), one abstracts away from the content of arguments and studies their external relations together with the acceptability conditions that constitute sets of justified arguments. In structured argumentation, the content of arguments—including claims, reasons, argument schemes, and premises—is additionally taken into account, which, among others, yields various types of attacks (Modgil and Prakken, 2014; Pollock, 1987; Walton and Reed, 2003) (cf. page 217).

[13] We do not consider argument strengths. Thus, every 'attack' between arguments is taken as a defeat, i.e., a successful attack. See (Modgil and Prakken, 2013) for a discussion of argument strength and defeat.

[14] We point out that the absence of weakening as a structural rule of DAC guarantees that all DAC-arguments triggered by a knowledge base $\mathcal{K}$ are void of free-floaters (cf. Remark 6.1).

Definition 6.10 (Stable Semantics and Nonmonotonic Inference)**.** *Let $\langle \mathsf{Arg}, \mathsf{Att}\rangle$ be an $\mathcal{AF}$ and let $\mathcal{E} \subseteq \mathsf{Arg}$:*

- $\mathcal{E}$ defeats *an argument $a \in \mathsf{Arg}$ if there is a $b \in \mathcal{E}$ that defeats a, i.e., $(b, a) \in \mathsf{Att}$;*
- $\mathcal{E}$ *is* conflict-free *if it does not defeat any of its own elements;*
- $\mathcal{E}$ *is* stable *if it is conflict-free and defeats all $b \in \mathsf{Arg} \setminus \mathcal{E}$.*

Let `Stable` *be the set of stable extensions of $\mathcal{AF}$. We define skeptic (s), skeptic* (s^*), and credulous (c) nonmonotonic inference as follows:*

- $\mathcal{AF} \mid\!\sim^{s}_{\text{stable}} \varphi$ *iff for each $\mathcal{E} \in$* `Stable`*, there is an $a \in \mathcal{E}$ concluding φ;*
- $\mathcal{AF} \mid\!\sim^{s^*}_{\text{stable}} \varphi$ *iff there is an $a \in \bigcap$* `Stable` *concluding φ;*
- $\mathcal{AF} \mid\!\sim^{c}_{\text{stable}} \varphi$ *iff there is an $\mathcal{E} \in$* `Stable` *s.t. there is an $a \in \mathcal{E}$ concluding φ.*

The use of DAC-arguments introduces nuances in argumentative inference. This is reflected in the distinction between skeptic and skeptic* inference: the consequence relation $\mid\!\sim^{s}$ denotes a conclusion shared by all stable extensions, whereas $\mid\!\sim^{s^*}$ denotes a shared argument by all stable extensions. In the context of DAC explanations, we can speak of *shared reasons* with respect to $\mid\!\sim^{s^*}$. The distinction between $\mid\!\sim^{s}$ and $\mid\!\sim^{s^*}$ also relates to our discussion of floating conclusions (see page 219). Last, $\mid\!\sim^{c}$ denotes the existence of reasons in favor of a conclusion φ with respect to some stable extension.

Example 6.7 (Example 6.1 cont.)**.** *The $\mathcal{AF}$ in Figure 6.1 is defined by $\mathsf{Arg} = \{a, b, c, d\}$ and $\mathsf{Att} = \{(b, c), (b, d)\}$. It has exactly one stable extension $\{a, b\}$. Consequently, the three inference relations equate: $\mathcal{AF} \mid\!\sim^{s,s^*,c}_{\text{stable}} (\neg t)^o$. In other words, given that Billy does not go to help her neighbors, she ought not to tell them she is coming. This is the desired outcome of this CTD scenario.*

Example 6.8 (Example 6.3 cont.)**.** *Joan is faced with a dilemma of conflicting duties. The $\mathcal{AF}$ of Figure 6.2 represents this conflict, where $\mathsf{Arg} = \{a, b, c_1, c_2, d, e\}$ and $\mathsf{Att} = \{(e, a), (e, c_1), (e, d), (d, e), (d, c_2), (d, b)\}$. The $\mathcal{AF}$ has two* stable *extensions $\{a, c_1, d\}$ and $\{b, c_2, e\}$. The extensions defend the views that Joan ought to return the hammer, respectively prevent harm. Hence, $\mathcal{AF} \mid\!\sim^{c}_{\text{stable}} r^o$, $\mathcal{AF} \mid\!\sim^{c}_{\text{stable}} p^o$, whereas $\mathcal{AF} \not\mid\!\sim^{c}_{\text{stable}} (r \wedge p)^o$, $\mathcal{AF} \not\mid\!\sim^{s}_{\text{stable}} r^o$, and $\mathcal{AF} \not\mid\!\sim^{s}_{\text{stable}} p^o$. For the floating conclusion that Joan ought to either return the hammer or prevent harm from being done $(r \vee p)^o$, we have $\mathcal{AF} \mid\!\sim^{s}_{\text{stable}} (r \vee p)^o$ but $\mathcal{AF} \not\mid\!\sim^{s^*}_{\text{stable}} (r \vee p)^o$.*

To illustrate the utility of our approach with respect to existing explanation techniques in formal argumentation (Objective 4), we consider the notion of related admissibility by Fan and Toni (2015a).

Definition 6.11 (Related Admissibility (Fan and Toni, 2015a))**.** *Let $\langle \mathsf{Arg}, \mathsf{Att} \rangle$ be an AF and let $\mathcal{E} \subseteq \mathsf{Arg}$:*

- *An extension $\mathcal{E}$ is* admissible *if it is conflict-free and $\mathcal{E}$ defeats all arguments defeating some $a \in \mathcal{E}$;*
- *An argument a* defends *b iff $a = b$, or there is a c such that a defeats c and c defeats b, or there is a c such that a defends c and c defends b.*
- *A set $\mathcal{E}_a \subseteq \mathsf{Arg}$ is* related admissible with topic *a iff $a \in \mathcal{E}_a$, for all $b \in \mathcal{E}_a$, b defends a, and $\mathcal{E}_a$ is admissible.*

In other words, a related admissible set $\mathcal{E}_a$ identifies the *relevant* arguments that justify the acceptability of a. Let $\mathcal{E}^+ = \{a \in \mathsf{Arg} \mid \mathcal{E}$ defeats $a\}$ and $\mathcal{E}^- = \{a \in \mathsf{Arg} \mid a$ defeats some $b \in \mathcal{E}\}$ be the set containing the arguments that are defeated by $\mathcal{E}$, respectively defeat arguments in $\mathcal{E}$. Using the above, we can explain why certain obligations hold.

Example 6.9 (Example 6.3 cont.)**.** *Recall that, in Example 6.3, argument b concludes that Joan is obliged to prevent harm. The answer to "why is Joan obliged to prevent harm?" is given by the related admissible set $\mathcal{E}_b = \{b, e\}$. Furthermore, the sets $\mathcal{E}_b^- = \{d\}$ and $\{d\}^- \cap \mathcal{E}_b = \{e\}$ explain that the only counterargument to b is argument d, the latter which is defeated by argument e expressing that the norm $(\top, r)$ used in d is inapplicable given the reasons $(\top, p)$ and $\neg(r \wedge p)^c$ offered in e. Thus, Joan is obliged to prevent harm because of the applicable norm $(\top, p)$, together with the fact that, given she cannot both prevent harm and return Maxwell's hammer, the conflicting norm $(\top, r)$ is inapplicable. A similar explanation can be given in favor of Joan being obliged to return the hammer (that is, it remains a dilemma). We point out that the related admissible set $\mathcal{E}_b$ is a proper subset of the stable extension $\{b, c_2, e\}$ discussed in Example 6.8. It shows us that the argument c_2 is not relevant for explaining why Joan is obliged to prevent harm.*

The above shows us that undercutting defeats and DAC-induced argumentation frameworks enable a more refined analysis of the *relevant* norms explaining the (non-)acceptability of specific arguments and obligations. The DAC approach is, therefore, more precise compared to using maximally consistent sets of norms in the traditional I/O formalism.

6.6 Soundness and Completeness, Part 2

In this section, we provide the proofs for the soundness and completeness results between constrained I/O logics and DAC-instantiated argumentation frameworks (Theorem 6.2). With this, we accomplish Objective 3. The main theorem uses the following lemma, expressing that for each sequent whose constraints on the lhs are implied by the constraint set $\mathcal{C}$, there is a derivable sequent containing only members of $\mathcal{C}$ on the lhs.

Lemma 6.12. *If $\vdash_{\mathcal{S}} \Delta, \Gamma_1^c \Rightarrow \Sigma$ and $\mathcal{C} \vdash \bigwedge \Gamma_1$, then there is a $\Gamma_2 \subseteq \mathcal{C}$ for which $\vdash_{\mathcal{S}} \Delta, \Gamma_2^c \Rightarrow \Sigma$ and $\Gamma_2 \vdash \bigwedge \Gamma_1$.*

Proof. Since $\mathcal{C} \vdash \bigwedge \Gamma_1$, by the compactness of $\vdash$ there is a $\Gamma_2 \subseteq \mathcal{C}$ for which $\Gamma_2 \vdash \bigwedge \Gamma$. By **Ax** and R$\wedge$1, $\vdash_{\mathcal{S}} \Gamma_2^c \Rightarrow \gamma^c$ for each $\gamma \in \Gamma_2$. By $\vdash_{\mathcal{S}} \Delta, \Gamma_1^c \Rightarrow \Sigma$ and multiple applications of **Cut**, $\vdash_{\mathcal{S}} \Delta, \Gamma_2^c \Rightarrow \Sigma$. QED

Theorem 6.2 (Soundness and Completeness)**.** *Let $\mathcal{K} = \langle \mathcal{F}, \mathcal{N}, \mathcal{C} \rangle$ be a knowledge base. Let $\mathcal{R}$ be a set* deriv*-rules and $\mathcal{S}$ a set of* DAC*-rules according to Table 6.1.*

1. *If $\mathcal{N}' \in \mathsf{maxfam}_{\mathcal{R},\mathsf{L}}(\mathcal{K})$ then $\mathsf{Arg}(\mathcal{F}^f \cup \mathcal{N}' \cup \mathcal{C}^c)$ is a stable extension of $\mathcal{AF}_{\mathcal{S}}(\mathcal{K}) = \langle \mathsf{Arg}, \mathsf{Att} \rangle$.*

2. *If $\mathcal{A}$ is a stable extension of $\mathcal{AF}_{\mathcal{S}}(\mathcal{K}) = \langle \mathsf{Arg}, \mathsf{Att} \rangle$ then there is a $\mathcal{N}' \subseteq \mathcal{N}$ such that $\mathcal{N}' \in \mathsf{maxfam}_{\mathcal{R},\mathsf{L}}(\mathcal{K})$ for which $\mathcal{A} = \mathsf{Arg}(\mathcal{F}^f \cup \mathcal{N}' \cup \mathcal{C}^c)$.*

Proof. We prove both items consecutively.

Ad 1. Let $\mathcal{N}' \in \mathsf{maxfamily}_{\mathcal{R},\mathsf{L}}(\mathcal{K})$ and $\mathcal{A} = \mathsf{Arg}(\mathcal{F}^f \cup \mathcal{N}' \cup \mathcal{C}^c)$.

For conflict-freeness assume towards a contradiction that there are $a = \Delta^f, \Theta, \Gamma^c \Rightarrow \neg(\varphi, \psi) \in \mathcal{A}$ (where $\Theta \subseteq \mathcal{N}'$) and $b = \Omega, (\varphi, \psi) \Rightarrow \Sigma \in \mathcal{A}$ such that a attacks b. By Lemma 6.3 and since $(\varphi, \psi) \in \mathcal{N}'$, we have, $\Delta^f, \Theta, \Gamma^c, (\varphi, \psi) \Rightarrow \in \mathcal{A}$. There are two cases: $\Gamma^c = \emptyset$ or not. If $\Gamma^c = \emptyset$, by Lemma 6.7, $\Delta^f, \Theta, (\varphi, \psi) \Rightarrow \bot^o \in \mathcal{A}$. By Theorem 6.1, $\bot \in \mathsf{deriv}_{\mathcal{R},\mathsf{L}}(\Delta, \Theta \cup \{(\varphi, \psi)\})$. Since $\Theta \subseteq \mathcal{N}'$, this contradicts the $\mathcal{C}$-consistency of $\mathcal{N}'$. If $\Gamma^c \neq \emptyset$, by Lemma 6.7, $\Delta^f, \Theta, (\varphi, \psi) \Rightarrow \sigma^o \in \mathcal{A}$ for some σ for which $\sigma \vdash \neg \bigwedge \Gamma$. By Theorem 6.1, $\sigma \in \mathsf{deriv}_{\mathcal{R},\mathsf{L}}(\Delta, \Theta \cup \{(\varphi, \psi)\})$, which contradicts the $\mathcal{C}$-consistency of $\mathcal{N}'$.

For $\mathcal{A}$ defeats all $a \in \mathsf{Arg} \setminus \mathcal{A}$, let $a = \Delta_1^f, \Theta_1, \Gamma_1^c \Rightarrow \Sigma \in \mathsf{Arg} \setminus \mathcal{A}$, where $\Theta_1 \subseteq \mathcal{L}^n$. So, there is a $(\varphi, \psi) \in \Theta_1 \setminus \mathcal{N}'$. By the maximal consistency of $\mathcal{N}'$, $\mathcal{N}' \cup \{(\varphi, \psi)\}$ is inconsistent with $\mathcal{C}$. So, there is a $\theta \in \mathsf{deriv}_{\mathcal{R},\mathsf{L}}(\Delta_2, \Theta_2)$ for some $\Delta_2 \subseteq \mathcal{F}$ and $\Theta_2 \subseteq \mathcal{N}' \cup \{(\varphi, \psi)\}$ such that $\mathcal{C} \vdash \neg\theta$. By Theorem 6.1, $\Delta_2, \Theta_2 \Rightarrow \theta^o \in \mathsf{Arg}$. Note that $(\varphi, \psi) \in \Theta_2$ since otherwise $\Theta_2 \subseteq \mathcal{N}'$, which contradicts the consistency of $\mathcal{N}'$. By **R-C** and **R-N**, $\vdash_{\mathcal{S}} \Delta_2, \Theta_2 \setminus \{(\varphi, \psi)\}, (\neg\theta)^c \Rightarrow \neg(\varphi, \psi)$. By Lemma 6.12, $b = \Delta_2, \Theta_2 \setminus \{(\varphi, \psi)\}, \Gamma_2^c \Rightarrow \neg(\varphi, \psi) \in \mathsf{Arg}$ for some $\Gamma_2 \subseteq \mathcal{C}$ for which $\Gamma_2 \vdash \neg\theta$. Note that $b \in \mathcal{A}$ and b attacks a.

Ad 2. Let $\mathcal{A}$ be a stable extension of $\mathcal{AF}(\mathcal{K})$. Let $\mathcal{N}' = \{(\varphi, \psi) \in \mathcal{N} \mid \neg\exists a \in \mathcal{A} \text{ with } \mathsf{Con}(a) = \neg(\varphi, \psi)\}$. We first show that $\mathcal{A} = \mathsf{Arg}(\mathcal{F}^f \cup \mathcal{N}' \cup \mathcal{C}^c)$.

- Left-to-Right. Let $a \in \mathsf{Arg} \setminus \mathsf{Arg}(\mathcal{F}^f \cup \mathcal{N}' \cup \mathcal{C}^c)$ with $a = \Delta \Rightarrow \Gamma$. So, there is a $(\varphi, \psi) \in \Delta$ for which there is a $b \in \mathcal{A}$ with $b = \Theta \Rightarrow \neg(\varphi, \psi)$. So b attacks a and by the stability of $\mathcal{A}$, $a \notin \mathcal{A}$.
- Right-to-Left. Let $a \in \mathsf{Arg}(\mathcal{F}^f \cup \mathcal{N}' \cup \mathcal{C}^c)$. By the definition of $\mathcal{N}'$, there is no $b \in \mathcal{A}$ that attacks a and since $a \in \mathsf{Arg}$ and by the stability of $\mathcal{A}$, $a \in \mathcal{A}$.

It remains to show that $\mathcal{N}' \in \mathsf{maxfam}_{\mathcal{R},\mathsf{L}}(\mathcal{K})$, i.e., $\mathcal{N}'$ is (i) $\mathcal{C}$-consistent and (ii) maximal.

Ad (i). Assume towards a contradiction that $\mathcal{N}'$ is inconsistent with $\mathcal{C}$. So, there is a $\theta \in \mathsf{deriv}_{\mathcal{R},\mathsf{L}}(\Delta, \Theta)$ for some $\Delta \subseteq \mathcal{F}$ and $\Theta \subseteq \mathcal{N}'$ for which $\mathcal{C} \vdash \neg\theta$. By Theorem 6.1, $a = \Delta^f, \Theta \Rightarrow \theta^o \in \mathcal{A}$.

We first show that $\Theta \neq \emptyset$. Assume towards a contradiction that $\Theta = \emptyset$. If $\mathbf{TP} \notin \mathcal{S}$, by Lemma 6.5.3, $\vdash \theta$ and thus $\mathcal{C}$ is inconsistent which is a contradiction. If $\mathbf{TP} \in \mathcal{S}$ then, by Lemma 6.5.2, $\Delta \vdash \theta$. However, then $\mathcal{F} \cup \mathcal{C}$ is inconsistent, which is a contradiction. Thus, $\Theta \neq \emptyset$.

Let $(\varphi, \psi) \in \Theta$. By **R-N** and **R-C**, $\vdash_{\mathcal{S}} \Delta, \Theta \setminus \{(\varphi, \psi)\}, (\neg\theta)^c \Rightarrow \neg(\varphi, \psi)$. By Lemma 6.12, there is a $\Gamma \subseteq \mathcal{C}$ for which $b = \Delta, \Theta \setminus \{(\varphi, \psi)\}, \Gamma^c \Rightarrow \neg(\varphi, \psi) \in \mathcal{A}$. Since b attacks a, we reached a contradiction to the conflict-freeness of $\mathcal{A}$. Altogether this shows that $\mathcal{N}'$ is consistent with $\mathcal{C}$.

Ad (ii). Assume for a contradiction that there is a $(\varphi, \psi) \in \mathcal{N} \setminus \mathcal{N}'$ such that $\mathcal{N}' \cup \{(\varphi, \psi)\}$ is consistent with $\mathcal{C}$, i.e., $\mathcal{N}'$ is not maximal. By the definition of $\mathcal{N}'$, there is a $b = \Delta^f, \Theta, \Gamma^c \Rightarrow \neg(\varphi, \psi) \in \mathcal{A}$. By Lemma 6.3, $\vdash_{\mathcal{S}} \Delta^f, \Theta, (\varphi, \psi), \Gamma^c \Rightarrow$. By Lemma 6.7, $\vdash_{\mathcal{S}} \Delta^f, \Theta, (\varphi, \psi) \Rightarrow \sigma^o$ such that $\sigma \vdash \neg \bigwedge \Gamma$. By Theorem 6.1, $\sigma \in \mathsf{deriv}_{\mathcal{R},\mathsf{L}}(\Delta, \Theta \cup \{(\varphi, \psi)\})$ which shows that $\mathcal{N}' \cup \{(\varphi, \psi)\}$ is inconsistent with $\mathcal{C}$ (note that $\Gamma \subseteq \mathcal{C}$). Contradiction. This completes our proof.

QED

6.7 Reasoning About Relevance

In this section, we further address Objective 4 and consider an extension of DAC with relevance rules. Concerning relevance, it can easily be observed that DAC suffers from the same problem as deriv does (cf. Remark 6.2 on page 222). Namely, DAC-arguments $\Delta \Rightarrow \Gamma$ can be derived where Δ contains norms and facts that are not strictly relevant for the conclusion expressed in Γ. Relevance rules yield more concise arguments, eliminating those arguments containing irrelevant reasons. The result is a more explanatory DAC-induced argumentation framework. We start with an example.

Example 6.10 (Irrelevant Reasons). *Let $\mathcal{F} = \{p\}$ and $\mathcal{N} = \{(p,q),(p,r)\}$ be the knowledge base. Consider the following* DAC*-derivation:*

$$\dfrac{\dfrac{}{p^f,(p,q) \Rightarrow q^o}\,\textbf{\textit{Detach}} \qquad \dfrac{\dfrac{}{p^f,(p,r) \Rightarrow r^o}\,\textbf{\textit{Detach}} \qquad \dfrac{}{q^o, r^o \Rightarrow q^o}\,\textbf{\textit{Ax}}}{q^o, p^f, (p,r) \Rightarrow q^o}\,\textbf{\textit{Cut}}}{p^f, (p,r), (p,q) \Rightarrow q^o}\,\textbf{\textit{Cut}}$$

Clearly, the norm (p,r) is irrelevant for concluding the obligation q^o. In fact, the following derivation contains sufficient reasons for concluding q^o:

$$\dfrac{}{p^f,(p,q) \Rightarrow q^o}\,\textbf{\textit{Detach}}$$

We extend DAC with *relevance rules*. Such rules exclude arguments like $p^f, (p,r), (p,q) \Rightarrow q^o$ from Example 6.10 as irrelevant. The basic idea is that these rules conclude which sequents are irrelevant by *eliminating* them. We use the sequent arrow $\not\Rightarrow$ to denote that the sequent is eliminated. Thus, an argument $\Delta \not\Rightarrow \Theta$ is interpreted as "the argument $\Delta \Rightarrow \Theta$ is eliminated since it contains irrelevant reasons".

We consider the following general relevance rule:

$$\dfrac{\Delta, \Delta' \Rightarrow \Theta \qquad \Delta \Rightarrow \Theta}{\Delta, \Delta' \not\Rightarrow \Theta}\ \mathbf{Rel}^a$$

where the side-condition (a) requires that $\Delta' \neq \emptyset$.[15] That is $\Delta \Rightarrow \Theta$ must be strictly more relevant. **Rel** ensures that only arguments with minimal support are allowed. The rule subsumes the following rule, which eliminates irrelevant defeating arguments:

$$\dfrac{\Delta, \Delta' \Rightarrow \neg(\varphi, \psi) \qquad \Delta, (\varphi, \psi) \Rightarrow}{\Delta, \Delta' \not\Rightarrow \neg(\varphi, \psi)}\ \mathbf{Rel'}^a$$

with the side-condition (a) $\Delta' \neq \emptyset$. The difference between the two rules is that the former additionally eliminates sequents with irrelevant facts and constraints.

In relation to Example 6.10, we obtain the following derivation:

$$\dfrac{\dfrac{\vdots}{p^f, (p,r), (p,q) \Rightarrow q^o} \qquad \dfrac{}{p^f, (p,q) \Rightarrow q^o}\ \mathbf{Detach}}{p^f, (p,r), (p,q) \not\Rightarrow q^o}\ \mathbf{Rel}$$

Definition 6.12 (Irrelevant Arguments)**.** *Let* $\Delta \Rightarrow \Gamma$ *be a* DAC*-derivable argument. We say that* $\Delta \Rightarrow \Gamma$ *is an* irrelevant *argument whenever there is a* DAC*-derivation applying* ***Rel*** *deriving* $\Delta \not\Rightarrow \Gamma$*. We say the argument is* relevant *otherwise.*

There are two ways to employ irrelevant arguments of the form $p^f, (p,r), (p,q) \not\Rightarrow q^o$ in our argumentative characterization of nonmonotonic normative reasoning:

1. We exclude irrelevant arguments from the DAC-instantiated $\mathcal{AF}$s;
2. We include irrelevant arguments in the DAC-instantiated $\mathcal{AF}$s.

The upshot of the first approach is that the DAC-instantiated $\mathcal{AF}$s are smaller, i.e., containing fewer and more concise arguments. The upshot of the second approach is that we can use irrelevant arguments to explain why specific reasons are irrelevant for explaining an obligation. For instance, in the context of dialogues an explainee may

[15] Recall that sequents are expressed in terms of finite regular sets.

provide irrelevant information (such as irrelevant counterarguments) while querying for a deontic explanation. Then, the explainer may provide a more relevant argument attacking it, thus pointing out its irrelevance. In what follows, we only pursue the first option and leave the second for future work (see Section 6.8).

Remark 6.4. *We address a possible objection to the two rules* ***Rel'*** *and* ***Rel****. Consider the knowledge base* $\mathcal{F} = \{p \wedge q\}$*,* $\mathcal{N} = \{(p, r)\}$*,* $\mathcal{C} = \emptyset$*. The following two arguments are* DAC*-derivable:* $a = p^f, (p, r) \Rightarrow r^o$ *and* $b = (p \wedge q)^f, (p, r) \Rightarrow r^o$*. Clearly, argument* a *is more relevant than argument* b *since argument* b *contains the redundant fact* q *as part of the conjunction* $p \wedge q$*. The rule* ***Rel*** *is insufficient for concluding* $b' = (p \wedge q)^f, (p, r) \not\Rightarrow r^o$*. However, this is a desired effect. Namely, in Definition 6.9 we defined* DAC*-instantiated* $\mathcal{AF}$*s to contain only those arguments* $\Delta \Rightarrow \Theta$ *triggered by the knowledge base* $\mathcal{K} = \langle \mathcal{F}, \mathcal{N}, \mathcal{C} \rangle$*, that is, where* $\Delta \subseteq \mathcal{F} \cup \mathcal{N} \cup \mathcal{C}$*. Since we do not require knowledge bases to be either minimal or elementary (think of the set* $\mathcal{F}$ *decomposed to literals only), we want to ensure that only those relevant arguments* triggered by $\mathcal{K}$ *are considered. The argument* a *is in fact not part of the* DAC*-instantiated* $\mathcal{AF}$*, whereas* b *is. The rule* ***Rel*** *ensures that this is the case.*

Since **Rel** subsumes the rule **Rel'**, we only consider extensions of DAC with the former (i.e., **Rel'** is admissible in the light of **Rel**).

Definition 6.13 (Relevance-aware Deontic Argumentation Calculi)**.** *Let* $\mathsf{DAC}_{\mathcal{S}}$ *be a system from Definition 6.8. We write* $\mathsf{DAC}^{\mathsf{r}}_{\mathcal{S}}$ *to denote* $\mathsf{DAC}_{\mathcal{S}}$ *extended with the rule* ***Rel****. We write* ***Rel***$\vdash_{\mathcal{S}} \Delta \Rrightarrow \Gamma$ *if* $\Delta \Rrightarrow \Gamma$ *is* $\mathsf{DAC}^{\mathsf{r}}_{\mathcal{S}}$*-derivable (with* $\Rrightarrow \in \{\Rightarrow, \not\Rightarrow\}$*).*

We observe that **Rel** can only be applied as the last rule of a derivation. The reason is that no rule in $\mathsf{DAC}^{\mathsf{r}}_{\mathcal{S}}$ takes irrelevant arguments as its premise. For now, this suffices since our interest is not in DAC-reasoning with irrelevant arguments but in obtaining more concise and explanatory DAC-induced argumentation frameworks.[16] Consequently, since all lemmas proven in the context of DAC are about $\Rightarrow$ sequents, they also hold in the context of $\mathsf{DAC}^{\mathsf{r}}$. The following lemma demonstrates height-preserving derivability in the light of **Rel**.

Lemma 6.13. *Let* $\mathsf{DAC}_{\mathcal{S}}$ *be a system from Definition 6.8 and let* $\mathsf{DAC}^{\mathsf{r}}_{\mathcal{S}}$ *be a corresponding system from Definition 6.13. Then,* $\vdash^{n}_{\mathcal{S}} \Delta \Rightarrow \Theta$ *iff* ***Rel***$\vdash^{n}_{\mathcal{S}} \Delta \Rightarrow \Theta$*.*

Proof. Left-to-Right. Straightforward since the rules of $\mathsf{DAC}_{\mathcal{S}}$ are rules of $\mathsf{DAC}^{\mathsf{r}}_{\mathcal{S}}$. Right-to-Left. The only rule in $\mathsf{DAC}^{\mathsf{r}}_{\mathcal{S}}$ different from those in $\mathsf{DAC}_{\mathcal{S}}$ is the **Rel** rule introducing sequents of the form $\Delta' \not\Rightarrow \Gamma'$. Hence, no other $\Delta' \Rightarrow \Gamma'$ sequent is additionally derivable by **Rel**. QED

[16] **Rel** causes irrelevant arguments to be permanently eliminated because no $\mathsf{DAC}^{\mathsf{r}}_{\mathcal{S}}$-rule enables reinstatement of a previously eliminated sequent $\Delta \not\Rightarrow \Gamma$. In Chapter 7, we propose an account of sequent-style reasoning with arguments that can be eliminated and, subsequently, reinstated.

Theorem 6.3. *Let* $\mathsf{DAC}^{\mathsf{r}}_{\mathcal{S}}$ *be a system from Definition 6.13. Let* $\Delta \subseteq \mathcal{L}$, $\psi \in \mathcal{L}$, *and* $\Theta \subseteq \mathcal{L}^n$. *Then,* $\textbf{\textit{Rel}} \vdash_{\mathcal{S}} \Delta^f, \Theta \Rightarrow \psi^o$ *iff* $\psi \in \mathsf{deriv}_{\mathcal{R},\mathsf{L}}(\Delta, \Theta)$.

Proof. Since we only consider $\Delta^f, \Theta \Rightarrow \psi^o$ arguments, the result follows immediately from Theorem 6.1 and Lemma 6.13. QED

We prove that $\mathsf{DAC}^{\mathsf{r}}$-instantiated $\mathcal{AF}$s—excluding irrelevant arguments—preserve soundness and completeness with respect to the considered class of constrained I/O logics.

Definition 6.14 ($\mathsf{DAC}^{\mathsf{r}}$-induced Argumentation Frameworks)**.** *Let* $\mathsf{DAC}^{\mathsf{r}}_{\mathcal{S}}$ *be a calculus and* $\mathcal{K} = \langle \mathcal{F}, \mathcal{N}, \mathcal{C} \rangle$ *a labelled knowledge base. We define a* $\mathsf{DAC}^{\mathsf{r}}$-induced argumentation framework $\mathcal{AF}^{\mathsf{r}}_{\mathcal{S}}(\mathcal{K}) = \langle \mathsf{Arg}^{\mathsf{r}}, \mathsf{Att}^{\mathsf{r}} \rangle$ *as follows:*

- $\Delta \Rightarrow \Gamma \in \mathsf{Arg}^{\mathsf{r}}$ *iff* $\Delta \Rightarrow \Gamma$ *is* $\mathsf{DAC}^{\mathsf{r}}_{\mathcal{S}}$*-derivable,* $\Delta \subseteq \mathcal{F} \cup \mathcal{N} \cup \mathcal{C}$, $\Gamma \subseteq \mathcal{L}^{io}$, *and* $\Delta \not\Rightarrow \Gamma$ *is not* $\mathsf{DAC}^{\mathsf{r}}_{\mathcal{S}}$*-derivable.*

Where $\mathsf{Att}^{\mathsf{r}}$ *is as defined as* Att *in Definition 6.9. We write* $\mathsf{Arg}^{\mathsf{r}}(\Sigma)$ *to denote the set of* $\mathsf{DAC}^{\mathsf{r}}_{\mathcal{S}}$*-derivable arguments* $\Delta \Rightarrow \Gamma$ *for which* $\Delta \subseteq \Sigma \subseteq \mathcal{L}^{io}$ *such that* $\Delta \not\Rightarrow \Gamma$ *is not* $\mathsf{DAC}^{\mathsf{r}}_{\mathcal{S}}$*-derivable.*

We first show a useful lemma that ensures that for every irrelevant argument triggered by the knowledge base, there exists a relevant argument that is a member of the argumentation framework based on that knowledge base.

Lemma 6.14. *Let* $\mathsf{DAC}^{\mathsf{r}}_{\mathcal{S}}$ *be a system from Definition 6.13 and* $\mathcal{K} = \langle \mathcal{F}, \mathcal{N}, \mathcal{C} \rangle$ *a knowledge base. Let* $\mathcal{AF}^{\mathsf{r}}_{\mathcal{S}}(\mathcal{K}) = \langle \mathsf{Arg}^{\mathsf{r}}, \mathsf{Att}^{\mathsf{r}} \rangle$ *be the* $\mathsf{DAC}^{\mathsf{r}}_{\mathcal{S}}$*-induced argumentation framework. For each* $\mathsf{DAC}^{\mathsf{r}}_{\mathcal{S}}$*-derivable* $\Delta \not\Rightarrow \Gamma$ *such that* $\Delta \subseteq \mathcal{F} \cup \mathcal{N} \cup \mathcal{C}$ *there is a* $\Delta' \Rightarrow \Gamma \in \mathsf{Arg}^{\mathsf{r}}$ *such that* $\Delta' \subset \Delta$ *and* $\Delta' \not\Rightarrow \Gamma$ *is* $\mathsf{DAC}^{\mathsf{r}}_{\mathcal{S}}$*-underivable.*

Proof. The proof is straightforward. First, observe that the side condition on **Rel** requires that for each irrelevant argument $\Delta \not\Rightarrow \Gamma$ there is a 'more' relevant argument $\Delta' \Rightarrow \Gamma$ that has strictly fewer reasons, i.e., $\Delta' \subset \Delta$. $\Delta' \Rightarrow \Gamma$, in turn, might also be derived as irrelevant, i.e., $\Delta' \not\Rightarrow \Gamma$, but since Δ is a finite set this process eventually halts (with the extreme case $\Rightarrow \Gamma$). QED

Theorem 6.4. *Let* $\mathcal{K} = \langle \mathcal{F}, \mathcal{N}, \mathcal{C} \rangle$ *be a knowledge base. Let* $\mathcal{R}$ *be a set of* deriv*-rules and* $\mathcal{S}$ *a set of corresponding* $\mathsf{DAC}^{\mathsf{r}}$*-rules according to Table 6.1.*

1. *If* $\mathcal{N}' \in \mathsf{maxfam}_{\mathcal{R},\mathsf{L}}(\mathcal{K})$ *then* $\mathsf{Arg}^{\mathsf{r}}(\mathcal{F}^f \cup \mathcal{N}' \cup \mathcal{C}^c)$ *is a stable extension of* $\mathcal{AF}^{\mathsf{r}}_{\mathcal{S}}(\mathcal{K}) = \langle \mathsf{Arg}, \mathsf{Att} \rangle$.
2. *If* $\mathcal{A}$ *is a stable extension of* $\mathcal{AF}^{\mathsf{r}}_{\mathcal{S}}(\mathcal{K}) = \langle \mathsf{Arg}^{\mathsf{r}}, \mathsf{Att} \rangle$ *then there is a* $\mathcal{N}' \subseteq \mathcal{N}$ *such that* $\mathcal{N}' \in \mathsf{maxfam}_{\mathcal{R},\mathsf{L}}(\mathcal{K})$ *for which* $\mathcal{A} = \mathsf{Arg}^{\mathsf{r}}(\mathcal{F}^f \cup \mathcal{N}' \cup \mathcal{C}^c)$.

Proof. We prove each item consecutively.

Ad 1. Let $\mathcal{N}' \in \mathsf{maxfam}_{\mathcal{R},\mathsf{L}}(\mathcal{K})$ and $\mathcal{A}^{\mathsf{r}} = \mathsf{Arg}^{\mathsf{r}}(\mathcal{F}^f \cup \mathcal{N}' \cup C^c)$.

First, we observe that $\mathcal{AF}^{\mathsf{r}}_{\mathcal{S}}(\mathcal{K}) = \langle \mathsf{Arg}^{\mathsf{r}}, \mathsf{Att}^{\mathsf{r}} \rangle$ is a sub-framework of $\mathcal{AF}_{\mathcal{S}}(\mathcal{K}) = \langle \mathsf{Arg}, \mathsf{Att} \rangle$ of Definition 6.9. In particular $\mathsf{Arg}^{\mathsf{r}} \subseteq \mathsf{Arg}$ and $\mathsf{Att}^{\mathsf{r}} \subseteq \mathsf{Att}$. Consequently, we know that $\mathcal{A}^{\mathsf{r}} \subseteq \mathsf{Arg}(\mathcal{F}^f \cup \mathcal{N}' \cup \mathcal{C}^c) = \mathcal{A}$. By Theorem 6.2 we know that $\mathcal{A}$ is a stable extension of $\mathcal{AF}_{\mathcal{S}}(\mathcal{K})$ and so $\mathcal{A}$ is conflict-free. Consequently, $\mathcal{A}^{\mathsf{r}}$ is conflict-free too.

We show that $\mathcal{A}^{\mathsf{r}}$ defeats all $a \in \mathsf{Arg}^{\mathsf{r}} \setminus \mathcal{A}^{\mathsf{r}}$. Let $a = \Delta_1^f, \Theta_1, \Gamma_1^c \Rightarrow \Sigma \in \mathsf{Arg}^{\mathsf{r}} \setminus \mathcal{A}^{\mathsf{r}}$ with $\Theta_1 \subseteq \mathcal{L}^n$. There is a $(\varphi, \psi) \in \Theta_1 \setminus \mathcal{N}'$ since if otherwise $\Theta_1 \subseteq \mathcal{N}'$ by the assumption $\mathcal{A}^{\mathsf{r}} = \mathsf{Arg}^{\mathsf{r}}(\mathcal{F}^f \cup \mathcal{N}' \cup C^c)$ we have $a \in \mathcal{A}^{\mathsf{r}}$, which contradicts our initial assumption. Consequently, $a \in \mathsf{Arg} \setminus \mathcal{A}$ too. Since $\mathcal{A}$ is stable there is a $b \in \mathcal{A}$ such that $b = \Delta_2^f, \Theta_2, \Gamma_1^c \Rightarrow \neg(\varphi, \psi)$, $\Delta_2^f \cup \Theta_2 \cup \Gamma^c \subseteq \mathcal{F}^f \cup \mathcal{N}' \cup \mathcal{C}^c$, and b attacks a. There are two options: if $b \in \mathsf{Arg}^{\mathsf{r}}$ (i.e., b is relevant), then since $\Delta_2^f \cup \Theta_2 \cup \Gamma^c \subseteq \mathcal{F}^f \cup \mathcal{N}' \cup \mathcal{C}^c$ we have $b \in \mathcal{A}^{\mathsf{r}}$ and we are done. If $b \notin \mathsf{Arg}^{\mathsf{r}}$ (i.e., b is irrelevant), then by Lemma 6.14 there is a relevant argument $b' \in \mathsf{Arg}^{\mathsf{r}}$ such that $b' = \Delta_3 \Rightarrow \neg(\varphi, \psi)$ and $\Delta_3 \subset \Delta_2^f \cup \Theta_2 \cup \Gamma^c$. Clearly, b' defeats a and $b' \in \mathcal{A}^{\mathsf{r}}$. We are done.

Ad 2. This proof is similar to the one of Theorem 6.2. The sole difference is that now we use Theorem 6.3 for the correspondence between deriv and $\mathsf{DAC}^{\mathsf{r}}$, instead of Theorem 6.1. QED

The upshot of the approach pursued in this section is that the argumentation framework contains only arguments with minimal support (Theorem 6.5). Thus, the argumentation is more concise and only generates arguments that provide minimal reasons for why certain norms are applicable or inapplicable.

Theorem 6.5. *Let $\mathsf{DAC}^{\mathsf{r}}_{\mathcal{S}}$ be a system from Definition 6.13 and $\mathcal{K} = \langle \mathcal{F}, \mathcal{N}, \mathcal{C} \rangle$ a knowledge base. Let $\mathcal{AF}^{\mathsf{r}}_{\mathcal{S}}(\mathcal{K}) = \langle \mathsf{Arg}^{\mathsf{r}}, \mathsf{Att}^{\mathsf{r}} \rangle$ be the $\mathsf{DAC}^{\mathsf{r}}_{\mathcal{S}}$ induced argumentation framework: if $\Delta \Rightarrow \Gamma \in \mathsf{Arg}^{\mathsf{r}}$, then $\Delta' \Rightarrow \Gamma \notin \mathsf{Arg}^{\mathsf{r}}$ for all $\Delta' \subset \Delta$.*

Proof. Suppose towards a contradiction that $\Delta \Rightarrow \Gamma, \Delta' \Rightarrow \Gamma \in \mathsf{Arg}^{\mathsf{r}}$ with $\Delta' \subset \Delta$. Since both arguments are $\mathsf{DAC}^{\mathsf{r}}_{\mathcal{S}}$-derivable, we can construct a derivation of both and apply **Rel**, thus deriving $\Delta' \not\Rightarrow \Gamma$. Since both $\Delta, \Delta' \subseteq \mathcal{K}$ by assumption, we have a contradiction with Definition 6.14. QED

6.8 Related Work and Future Research

In the remainder of this chapter, we discuss the related literature. We point out some open questions and future work directions along the way.

Default Logic. In Default Logic (Reiter, 1980), the main components are *default rules* and a set of *assumptions* (e.g., a partial description of the world). Default rules are defeasible, and the set of assumptions is taken as strict. A default rule is an inference rule of the form $(\varphi : \theta_1, \dots, \theta_n)/\psi$ where φ is the pre-requisite, $\theta_1, \dots, \theta_n$ signify the justification, and ψ is the conclusion. Roughly, such a rule is interpreted as "if φ holds, and there is no proof that $\theta_1, \dots, \theta_n$ do not hold, then ψ may be derived". A central aim is to preserve consistency while applying default rules. Default logic generates consistent extensions of inferred formulae utilizing a fixed point approach, e.g., see (Reiter, 1980; Straßer and Antonelli, 2019). Normal Defaults are defaults of the form $(\varphi : \psi)/\psi$, where the conclusion ψ may be inferred whenever the antecedent φ holds, and it is not derived that ψ does not hold. As shown by Parent (2011), under certain conditions prioritized Input/Output logics strongly relate to *greedy* reasoning with Reiter's (1980) normal default logic. The following remains to be investigated:

Open question 6.1. *Is there a formal correspondence between* DAC *and (variations of) Reiter's default logic?*

Variations of Reiter's default logic were developed by, e.g., Horty (2012) and Gelfond et al. (1991). In particular, Horty (1997) proposes Deontic Default Logic, which is strongly related to Input/Output logic (Parent, 2011) (see the discussion in Chapter 7, page 293).

Extensions of I/O logics. The I/O formalism knows other applications, including normative reasoning with consistency checks, permissions, and constitutive norms (Pigozzi and van der Torre, 2018; Tosatto et al., 2012). Due to the presence of norms as objects in the logical language of DAC, one can similarly introduce other types of norms, such as permissive and constitutive norms, together with logical rules defining their interaction. In particular, we aim to exploit the internalization of meta-reasoning in DAC for characterizing various types of permission (Makinson and van der Torre, 2003; Olszewski et al., 2021; Tosatto et al., 2012). For instance, negative permissions can be defined in terms of the absence of applicable norms to the contrary. DAC-rules that introduce negative permissions—denoted as objects of the form $(\varphi, \psi)_p$ with a corresponding detached formula ψ^p—would be of the following form:

$$\frac{}{\varphi, (\varphi, \psi)_p \Rightarrow \psi^p}\,\textbf{DetachPer} \qquad\qquad \frac{\Delta \Rightarrow \neg(\varphi, \neg\psi)}{\Delta \Rightarrow (\varphi, \psi)_p}\,\textbf{NegPer}$$

The rule on the left is a **Detach** rule for negative permission. The rule on the right derives a permissive norm $(\varphi, \psi)_p$ from the inapplicability of a regulative norm $(\varphi, \neg\psi)$ to the contrary.

Open question 6.2. *Does* DAC *extended with the rules* ***DetachPer*** *and* ***NegPer*** *characterize constrained I/O logics with negative permissions?*

Next, Bochman (2021) investigates I/O pairs as *production/explanation* rules, where the pairs (φ, ψ) and $(\theta.\varphi, \psi)$ are interpreted as "φ produces/explains ψ", respectively "After θ, φ produces/explains ψ". The logics do not satisfy identity (ID) and contain a falsity preservation rule $(\bot, \bot)$. Furthermore, they are shown to be conceptually close to the action description language C+ (Bochman, 2014; Giunchiglia et al., 2004). Adaptation of Bochman's (2021) work on abductive causal inference to the context of abductive deontic inference in DAC remains to be investigated.

Alternative proof systems for I/O logics. Lellmann (2021) proposes a sequent-style system for unconstrained I/O logics with consistency checks. It utilizes a translation from I/O to modal conditional logics, treating norms as dyadic modalities instead. The results hold for a class of I/O logics based on the systems $\mathcal{R}_1$ and $\mathcal{R}_3$ (defined on page 221). The calculi proposed by Lellmann (2021) are shown to possess the critical proof-theoretic property of cut-freeness and enjoy the property of analyticity (Negri et al., 2008) (see the footnote on page 10). Furthermore, decidability and complexity results are provided. Whether modifications of DAC enjoy similar properties remains to be determined. At the moment, DAC still depends on applications of **Cut** and the cut-like rule **L-CT**.

Open question 6.3. *Under what extensions of* DAC *can we ascertain cut-free calculi?*

Straßer et al. (2016) provide an adaptive logic approach to the I/O formalism using dynamic proof systems. First, they translate the monotonic unconstrained I/O logics into modal logics, using unary modalities to characterize 'input', 'output', and 'constraints'. Second, they use adaptive logics on top of these modal logics to characterize nonmonotonic constrained I/O logics (see also the discussion of adaptive logics in Chapter 7 on page 289). The class of I/O logics considered by Straßer et al. (2016) subsumes the class of logics discussed for DAC. For instance, they consider systems that contain the rule $(\bot, \bot)$ used in causal interpretations of I/O logics (Bochman, 2014). The use of modalities for 'input' and 'output' allows for reasoning about statements such as "input φ is not a reason for output ψ" or "input φ is a reason against output ψ". DAC uses labels instead of modalities, and we leave it for future work to investigate the expressive power of DAC in relation to these adaptive logics. Although the introduced adaptive logics are shown sound and complete for a large class of I/O logics, a correspondence with formal argumentation is not given. Straßer (2014) provides a general correspondence between adaptive logics and formal argumentation. Furthermore, the central purpose of DAC is to generate arguments that provide explicit reasons and facilitate explanations.

Last, Straßer and Arieli (2015) use sequent-based argumentation to model defeasible reasoning with Standard Deontic Logic. Relations to formalisms based on I/O logics are also discussed. Their presentation is not aimed at explanatory reasoning: e.g., the resulting sequents are not relevance-aware (the calculus has unrestricted weakening). Moreover, since norms are modeled with material implications, their approach allows for less fine-tuning of norms than in DAC. Furthermore, DAC additionally represents all standard constrained I/O systems.

Argumentative characterizations of normative reasoning. Dong et al. (2020), Governatori et al. (2018), Liao et al. (2018), and Straßer and Pardo (2021) study argumentative characterizations of normative systems employing priority orderings. The latter two use a version of the I/O system $\mathcal{R}_3$ extended with priorities. They all use languages restricted to literals, whereas our approach adopts a full propositional language. Liao et al. (2018) and Straßer and Pardo (2021) take arguments to consist only of (sets of) norms. The approach by Straßer and Pardo (2021) extends the work by Liao et al. (2018). Dong et al. (2020) and Governatori et al. (2018) additionally use deontic modalities to characterize obligations. It is left to future work to incorporate priority reasoning with a full propositional language in the more transparent context of DAC.

Open question 6.4. *How can we incorporate explicit priority reasoning in* DAC*?*

An alternative approach to modeling reasoning with norms is to instantiate ASPIC$^+$—as developed by Modgil and Prakken (2013; 2014)—with conditionals representing norms and a defeasible modus ponens rule. This approach leads to a "greedier" style of reasoning than our approach. To see this point, consider $\mathcal{F} = \emptyset$ and $\mathcal{N} = \{(\top, p), (p, q), (\top, \neg q)\}$. An ASPIC$^+$-based approach yields the obligation p with stable semantics since the argument for p from $(\top, p)$ is unchallenged. In contrast, our approach yields the argument $(\top, \neg q), (p, q) \Rightarrow \neg(\top, p)$ concluding the inapplicability of $(\top, p)$ (given **L-CT**). The latter is in line with the I/O approach to normative reasoning.

Multi-agent systems and the BOID architecture. The BOID architecture, as developed by Broersen et al. (2001), extends the Belief-Desire-Intention view on practical reasoning (Rao and Georgeff, 1995) with obligations. Obligations are taken as external motivational attitudes. In contrast to BDI logics (Rao and Georgeff, 1998), BOID is not a modal logic. The formalism processes input in order to output beliefs, obligations, intentions, and desires. Its main focus is on conflicts among beliefs, obligations, intentions, and desires. Conflicts are resolved by imposing priorities that indicate which of the four types overrules others in case of conflict. Broersen et al. (2001) propose several types of agents with different preferences (similar to the BDI formalism). To illustrate, a social agent prioritizes obligations over desires. The framework is in the spirit of Default Logic (Reiter, 1980) and Input/Output logic (Makinson and van der Torre, 2000). The input is a set of propositional formulae interpreted as the context. Beliefs, obligations, intentions, and desires are taken as conditionals similar to normal defaults and I/O pairs. Roughly, the context triggers conditionals, resulting in output, i.e., goals. Conflicts arise whenever an inconsistent set of goals is generated (Broersen et al., 2002). Priorities are employed to resolve conflicting output. Under different BOID formalisms, either single or multiple extensions are generated. Consistency constraints are employed to restrict the inference process while generating output (Broersen et al., 2002).

The BOID architecture shows a strong conceptual relation to the formalism developed in this chapter due to its similarity to Default Logic and Input/Output logic. Since BDI and BOID systems have been influential in the development of multi-agent systems in AI, a promising future research direction is the investigation of BOID in the context of DAC.

Open question 6.5. *Can we provide an argumentative characterization of BOID using extensions of* DAC*?*

Answering this question involves labeling different types of (φ, ψ) pairs to syntactically differentiate between beliefs, desires, intentions, and obligations. Furthermore, DAC must be extended with explicit priorities over these pairs.

Explanation and formal argumentation. A different type of explanation, not treated in this chapter, clarifies why certain norms are added to a normative code. Such explanations enhance our understanding of why normative codes are the way they are. Here, one may think of value-based reasoning where the inclusion of certain norms is explained by reference to the general values promoted by the norm (Bench-Capon, 2002; Bench-Capon, 2003; Bench-Capon and Sartor, 2003). To illustrate, one may answer a question of the form "why I am obliged to wear a safety belt while driving a car" by referring to "safety" as a value endorsed by society. We leave such explanations for future work.

In Section 6.5, we illustrated our approach using the semantic notion of *related admissibility*, developed by Fan and Toni (2015a). The notion identifies relevant information explaining the defensibility of an argument. Fan and Toni (2015a) extend their work to Assumption-Based Argumentation (Dung et al., 2009) for a more fine-grained analysis. Their definitions refer to assumptions (premises) and conclusions of structured arguments, which conceptually correspond to reasons (lhs) and conclusions (rhs) in DAC-arguments. We note that the relevance property that DAC enjoys (Section 6.7) makes the calculi particularly interesting for related-admissibility. Namely, relevant deontic arguments contain only relevant norms.

Borg and Bex (2021) provide an extension of the work by Fan and Toni (2015a) and study explanations of (non)-acceptance of arguments and formulae. For instance, a formula φ is not accepted if there is no derivable argument for it or all its supporting arguments are attacked. Consequently, an explanation of non-acceptance pinpoints the gaps in the knowledge base. Our use of adequate sequent calculi opens the door for reasoning about non-acceptance through *underivable sequents* denoted by $\nvdash \Delta \Rightarrow \Gamma$ (Bonatti and Olivetti, 2002). For instance, if $\nvdash \mathcal{F} \Rightarrow \varphi$ holds, we may also conclude that $\nvdash \mathcal{F}, (\varphi, \psi) \Rightarrow \psi^o$. In other words, if $\mathcal{F}$ does not imply φ, it is *a fortiori* insufficient to trigger the norm (φ, ψ). Also, see the work of Fan and Toni (2015b) and Saribatur et al. (2020) for other notions of non-acceptance.

The above works on explanation propose formalisms on top of abstract ad structured argumentation (see the footnote on page 240). These formalisms mainly deal with the relations between arguments, thus serving external explanation. Our approach accommodates explanation by constructing suitable arguments that facilitate internal and external explanations. Consequently, our approach focuses on a different level of explanation (the level of argument construction), and is, for that reason, compatible with these accounts. The work by Borg and Bex (2021), Fan and Toni (2015a), and Šešelja and Straßer (2013) can be readily applied to DAC-induced argumentation frameworks.

Dialogues and Irrelevance Attacks. Last, explanations typically occur in the context of dialogues, (Prakken, 2005; Walton, 2010). Dialogical episodes are often characterized by an exchange of reasons, questions, and (explanatory) arguments. Consequently, explanations are ideally tailored to the background of the explainee. This motivates the use of formal models of explanation based on dialogues (Arioua and Croitoru, 2015; Bex and Budzynska, 2012; Bex and Walton, 2016; Feldhus et al., 2022). There is literature on formal dialogues, ranging from inquiry, information-seeking, and persuasion dialogues of argumentation, to dialogues of deliberation (McBurney and Parsons, 2009). Several of these works provide dialogical generalizations of monological argumentation, e.g., see (Caminada, 2017). The seminal work by Prakken (2005) contains models for two-player persuasion dialogues that determine membership of the grounded semantics. Fan and Toni (2014) propose dialogue systems for Assumption-Based Argumentation dealing with three types of semantics. Furthermore, Modgil (2017) develops preference dialogues for admissible and grounded semantics (see Definition 7.11 in Chapter 7 for the definitions of grounded semantics). Closely related are argumentation games, which determine the acceptability of an argument with respect to a given semantics utilizing a restricted two-player dialogue (Modgil and Caminada, 2009).

In dialogues, agents often exchange arguments containing seemingly irrelevant information. In this respect, we can further exploit irrelevant DAC-arguments $\Delta \not\Rightarrow \Gamma$. In Section 6.7, we used irrelevance to obtain a more concise DAC-induced argumentation framework containing only arguments providing minimal support. There, we excluded irrelevant arguments from the argumentative discourse. Alternatively, we can include irrelevant arguments triggered by the knowledge base. Based on these arguments, 'irrelevance' attacks can be defined between arguments. Then, an argument $\Delta \Rightarrow \Gamma$ attacks an argument $\Delta' \Rightarrow \Gamma$ on the basis of being irrelevant, provided that $\Delta \subset \Delta'$ and $\Delta' \not\Rightarrow \Gamma$ is $\mathsf{DAC}^{\mathsf{r}}$-derivable. In the context of dialogues, this enables the explainee and explainer to pinpoint irrelevant premises to one another. We conjecture that the inclusion of irrelevant arguments and corresponding attacks preserves soundness and completeness.

Conjecture 6.1. *Soundness and completeness are preserved for nonmonotonic I/O logic and* $\mathsf{DAC}^{\mathsf{r}}$*-induced argumentation frameworks in the presence of explicit irrelevance attacks between arguments.*

* * *

In this chapter, we addressed *deontic explanations*, which are answers to why questions such as "Why is Joan obliged to return a borrowed weapon, despite her obligation to prevent harm?" (Example 6.3). We introduced a highly modular proof theoretic formalism called Deontic Argumentation Calculi (DAC), which explicitly formalizes reasons (Objective 1) and integrates meta-reasoning about the inapplicability of norms into the object language of proofs (Objective 2). We proved that DAC are sound and complete for a large class of monotonic Input/Output logics and employed DAC to provide a sound and complete argumentative characterization of the class of nonmonotonic constrained Input/Output logics (Objective 3). We discussed the explanatory nature of DAC-induced argumentation frameworks by applying the notion of related admissibility (Fan and Toni, 2015a) and by extending DAC with relevance rules that exclude irrelevant reasons (Objective 4). By identifying *relevant* norms explaining the (non-)acceptability of specific arguments and obligations, the DAC approach proved more precise compared to using maximally consistent sets of norms in the traditional I/O formalism.

CHAPTER 7

Defeasible Reasoning

In the previous chapter, we developed the monotonic sequent-style Deontic Argumentation Calculi (DAC). We showed that nonmonotonic inference of Input/Output (I/O) logics can be argumentatively characterized by argumentation frameworks instantiated with arguments generated by DAC. A natural question that arises from these results is whether DAC can be modularly extended with specific rules such that derivability in the resulting calculi corresponds directly to nonmonotonic inference. That is, can we provide a general proof-theoretic approach to defeasible normative reasoning? In this chapter, we provide several results addressing this question by pursuing a *more general objective*. We make use of two central observations.

First, an essential feature of nonmonotonic logics is that previous inferences can be retracted in the light of novel information (Chapter 1). For instance, in the context of normative reasoning (Chapter 6), one may find that an obligation must be revised in the context of exceptional circumstances. To illustrate, an agent may not be permitted to drive on the left side of the road. However, this prescription is revised in the *additional context* of overtaking another vehicle, where a permission to overtake via the left is triggered instead. Thus we say, in the context of nonmonotonic reasoning, the *status* of a formula as a logical conclusion may have to be *revised* (several times).

Second, since its beginning, formal argumentation (Dung, 1995) has proven to be a unifying framework for the representation of large classes of nonmonotonic logics (Arieli et al., 2021). In formal argumentation, inferences from a given knowledge base can be represented through *arguments*. The relations between arguments concluding conflicting information are defined in terms of *attacks* between these arguments. The resulting argumentation framework can then be analyzed to determine which arguments are collectively defensible (or acceptable). Various semantics have been proposed that define different types of defensible sets of arguments (Baroni et al., 2011). For many nonmonotonic logics, it can be shown that nonmonotonic inference of the logic corresponds to membership of specific semantic extensions (Arieli et al., 2021; Straßer, 2014) (e.g., autoepistemic logics, adaptive logics, default logics, and logic programs). The notion of attack is central in this respect since it expresses which arguments are compatible with one another and which are incompatible.

Our primary objective is to integrate these two central concepts of *revision* and *attack* into a proof-theoretic approach to nonmonotonic logic.

Objective 1. *Develop a class of proof systems for nonmonotonic reasoning with conflicting information where revision considerations based on attacks are fully integrated into the object level of the proofs.*

In this chapter, we introduce Annotated Calculi (AC, for short), which is a robust family of nonmonotonic sequent-style proof systems. Unlike in ordinary calculi, sequents derived in annotated calculi may still be retracted in the presence of conflicting sequents. Thus, inferences are made under stricter conditions. The calculi extend standard analytic sequent calculi (Gentzen, 1934; Negri et al., 2008) in the following way: First, we extend the language of sequents with *annotations* on sequents which represent the sequents' status in a derivation. It must be noted that unlike annotated logics (Abe et al., 2019), the annotations are attached to sequents instead of formulae in the language. Second, we extend these calculi with rules that represent various types of attacks between sequents, leading to a *revision* of the status (i.e., annotation) of the sequent under attack. The result is a novel proof-theoretic approach for nonmonotonic reasoning with conflicting information in which revision procedures are fully integrated on the object level of proofs.

An annotated sequent is an expression of the form $\Gamma \Rightarrow^{[\mathsf{a}]} \Delta$, where $\Gamma \Rightarrow \Delta$ is an ordinary sequent and the superscript [a] is the annotation of the sequent (see Chapter 6, page 224 for an introduction to sequent systems). We use the following annotations: The annotation [i] means that the sequent is introduced (conditionally accepted) but is not yet inferred (finally accepted) because a counter-sequent may still attack it. The annotation [e] means that the sequent is eliminated because of an attack by an accepted sequent. Last, the annotation [!] is attached to finally accepted sequents, whose attackers are counter-attacked altogether.

To reach the first objective, we must ensure that the resulting proof systems behave well. As a minimal requirement, we first demonstrate that the inference relation of AC is nonmonotonic. Second, we show that in the absence of conflict (i.e., the presence of a consistent knowledge base), the nonmonotonic inference relation of AC corresponds to the monotonic inference relation of the underlying base logic. Last, we show that the resulting nonmonotonic logics are *paraconsistent*, i.e., given an inconsistent knowledge base, AC does not generate an explosive set of finally accepted conclusions.

In nonmonotonic reasoning, conflicting information often yields several alternative extensions of a theory. There are two main approaches to drawing conclusions from such extensions. *Skeptic* inference allows us only to infer what is common to all of these extensions. *Credulous* inference considers each extension as an acceptable solution, and we may infer a formula whenever it belongs to one such extension (even though this may contradict with certain inferences from other extensions). In the context of AC, we interpret sequents annotated with [!] as skeptically inferred and those with [i] as credulously inferred. An important question that must be answered is whether there is a correspondence between various annotations of sequents and skeptical and credulous inference in formal argumentation.

Objective 2. *Determine the formal correspondence between the nonmonotonic inference relations of annotated calculi and formal argumentation.*

Coming back to our initial aim, we will leverage the results for AC to enhance the *Deontic Argumentation Calculi* (DAC) of Chapter 6 to obtain a class of nonmonotonic proof systems for normative reasoning. In particular, since DAC was shown to correspond to a large class of I/O logics, we require that the resulting calculi will preserve a similar correspondence. This chapter contains the first step towards this result by addressing the following objective.

Objective 3. *Extend* DAC *with rules such that derivability in the resulting calculi characterizes skeptical and credulous nonmonotonic inference in formal argumentation.*

In addressing this objective, we provide a class of sequent calculi for defeasible normative reasoning based on the I/O formalism.

Contributions. A distinctive property of annotated sequent calculi is their modularity. Namely: they can be based on any (propositional) logic with a sound and complete sequent calculus and any set of attack rules. In this chapter, we demonstrate this novel approach, initially introduced in (Arieli et al., 2022a), by focusing on *defeat* attacks. Other attack rules can be found in Arieli and Straßer, 2019. We adopt defeat attacks mainly due to their commonness and simplicity (Baroni et al., 2018; Toulmin, 1958). The technical contributions are threefold:

- First, we show that the resulting proof systems are nonmonotonic and paraconsistent. The latter expresses that a contradictory set of premises does not have an explosive set of finally accepted conclusions.
- Second, we demonstrate that the derivations of annotated calculi faithfully represent the semantics of logical argumentation frameworks.
- Third, reasoning aspects that are typically reserved for the meta-level of nonmonotonic reasoning (i.e., credulous and skeptical inference) are fully internalized in annotated sequent calculi and can be expressed in the object-level language of the derivation.

This work is the first fully integrate both skeptical and credulous inference of logical argumentation through a proof-theoretic approach. From a conceptual point of view, our approach demonstrates that, despite the calculi's relative simplicity, their derivations are particularly appropriate for modeling and describing inference processes involving the revision of beliefs and defeasible reasoning.

The majority of the results in this chapter are proof-theoretical. Additionally, we extend our formalism to the context of defeasible normative reasoning and illustrate the obtained calculi by analyzing deontic benchmark examples.

Differences. We first published the results in this chapter in (Arieli et al., 2022a). Novel contributions presented in this chapter are Proposition 7.1 (Section 7.1), Proposition 7.3 and 7.7 (Section 7.3), and the entire Section 7.5 on the integration of Deontic Argument Calculi (Chapter 6) with annotated calculi. Furthermore, we also introduce the novel annotation $[\bot]$ representing finally eliminated sequents in Section 7.5.

Outline. In Section 7.1, we introduce *Annotated Calculi* (AC). To illustrate our approach, we discuss various examples in Section 7.2. After that, in Section 7.3, we demonstrate various properties of the AC-inference relation. The correspondence to logical argumentation is addressed in Section 7.4. In Section 7.5, we extend Deontic Argumentation Calculi DAC with AC-rules and annotations and prove correspondence with formal argumentation. Last, we discuss related and future work in Section 7.6.

7.1 Annotated Sequent Calculi

We aim to develop a nonmonotonic proof-theoretic approach that is modular for a large class of underlying base logics (Objective 1). Therefore, we provide general definitions of the languages and logics forming this class.

Definition 7.1 (Base Language $\mathcal{L}$). *Let* Atoms *be a set of propositional atoms* $p, q, r, \dots$ *(possibly indexed). Let* $\mathcal{L}$ *be an arbitrary well-formed propositional language recursively defined over* Atoms *and a set of propositional connectives* $\mathsf{Connectives}$*. We require that* $\mathcal{L}$ *contains at least a negation and a conjunction operator, i.e.,* $\neg, \wedge \in \mathsf{Connectives}$.

We use lowercase Greek letters $\varphi, \psi, \chi, \dots$ to refer to arbitrary formulae of $\mathcal{L}$. Arbitrary sets of formulae from $\mathcal{L}$ are denoted by $\mathcal{S}, \mathcal{T}$ and we use upper case Greek letters $\Gamma, \Delta, \Sigma, \dots$ to refer to finite sets of formulae. Both formulae and sets may be indexed.

Definition 7.2 (Base Logic L). *Let* $\mathsf{L} = \langle \mathcal{L}, \vdash \rangle$ *be a* propositional logic*, where* $\mathcal{L}$ *is a propositional language from Definition 7.1 and* $\vdash_{\mathsf{L}}$ *is a consequence relation on* $\wp(\mathcal{L}) \times \mathcal{L}$ *(henceforth, we omit the subscript* L*). Let* $\mathcal{S} \subseteq \mathcal{L}$*,* $\vdash$ *satisfies the following properties:* reflexivity*,* monotonicity*,* transitivity*,* non-triviality*,* structurality*, and* compactness *(see Definition 6.2 for the formal specifications of these properties). We assume that negation* $\neg$ *and conjunction* $\wedge$ *satisfy the following* $\vdash$*-properties:*

- $p \nvdash \neg p$ *and* $\neg p \nvdash p$ *(for* $p \in \mathsf{Atoms}$*);*

- *$\mathcal{S} \vdash \psi \wedge \varphi$ iff $\mathcal{S} \vdash \psi$ and $\mathcal{S} \vdash \varphi$.*

As a consequence of the above characterization of $\wedge$, we have that $\varphi \wedge \psi \vdash \varphi$, $\varphi \wedge \psi \vdash \psi$, and $\varphi, \psi \vdash \varphi \wedge \psi$. Thus, L also satisfies $\mathcal{S}, \varphi, \psi \vdash \sigma$ iff $\mathcal{S}, \varphi \wedge \psi \vdash \sigma$. In the sequel, we write $\bigwedge\Gamma$ for the conjunction of all formulae in Γ.

We are interested in defining annotated *sequent-style* proof systems (Gentzen, 1934). For this reason, we limit ourselves to base logics L that have a corresponding sound and complete sequent calculus LC. We refer to Chapter 6, page 224, for a discussion of sequent calculi.

Definition 7.3 (Base Calculus LC)**.** *Let $\mathcal{L}$ be a language from Definition 7.1. A sequent based on $\mathcal{S} \subseteq \mathcal{L}$ is a structure of the form*

$$\Gamma \Rightarrow \Delta$$

where $\Rightarrow$ is a symbol not in $\mathcal{L}$, $\Gamma \subseteq \mathcal{S}$, and $\Delta \subseteq \mathcal{L}$ is either the empty-set or a singleton-set.

Let L be a logic from Definition 7.2, $\Gamma \subseteq \mathcal{L}$, and $\Delta \subseteq \mathcal{L}$ be either the empty-set or a singleton-set. We say that a sequent calculus LC is sound and complete with respect to L whenever $\Gamma \Rightarrow \Delta$ is LC-derivable iff $\Gamma \vdash \Delta$.

The restriction imposed on Δ in the above definition facilitates a large class of base logics. As a terminological clarification, for an arbitrary sequent rule of the form

$$\frac{\Gamma_1 \Rightarrow \Delta_1 \qquad \cdots \qquad \Gamma_n \Rightarrow \Delta_n}{\Gamma_m \Rightarrow \Delta_m}$$

we refer to the topmost sequents $\Gamma_1 \Rightarrow \Delta_1, \ldots, \Gamma_n \Rightarrow \Delta_n$ as the rule's *conditions* and refer to bottom sequent $\Gamma_m \Rightarrow \Delta_m$ as the rule's *conclusion*. Last, we refer to the left-hand side (lhs) Γ_i as the sequent's premises and to the right-hand side (rhs) Δ_i as the sequent's conclusion.

Definition 7.4 (Annotated Sequents)**.** *Let LC be a sequent calculus from Definition 7.3. An* annotated LC-sequent *(annotated sequent, for short) is a structure of the form $\Gamma \Rightarrow^{[\mathsf{a}]} \Delta$, where $\Gamma \Rightarrow \Delta$ is an LC-sequent and $\mathsf{a} \in \{\mathsf{i}, \mathsf{e}, !\}$. We denote by $s[\mathsf{a}]$ the sequent s whose annotation is $[\mathsf{a}]$ with $\mathsf{a} \in \{\mathsf{i}, \mathsf{e}, !\}$. We use $[*]$ to express that the annotation of the sequent is arbitrary.*

Recall that the annotated sequent $\Gamma \Rightarrow^{[\mathsf{i}]} \Delta$ denotes an *introduced* sequent (conditionally accepted), $\Gamma \Rightarrow^{[\mathsf{e}]} \Delta$ expresses that the sequent is *eliminated* (conditionally refuted), and $\Gamma \Rightarrow^{[!]} \Delta$ denotes a finally accepted sequent.

Next, we define Annotated Calculi (AC, for short), extending a sequent calculus LC of the base logic L with *annotation revision rules*. The latter rules are presented in Figure 7.1. We briefly discuss the intuition behind these rules.

$$\frac{\Gamma_1, \Gamma_1' \Rightarrow^{[\mathsf{i}]} \Delta \qquad \Gamma_2 \Rightarrow^{[\mathsf{i}]} \psi_2 \qquad \psi_2 \Rightarrow^{[*]} \neg \bigwedge \Gamma_1}{\Gamma_1, \Gamma_1' \Rightarrow^{[\mathsf{e}]} \psi_1}\ \mathbf{Def}^a$$

$$\frac{\Gamma_1, \Gamma_1' \Rightarrow^{[\mathsf{e}]} \Delta \qquad \Gamma_2 \Rightarrow^{[\mathsf{e}]} \psi_2 \qquad \psi_2 \Rightarrow^{[*]} \neg \bigwedge \Gamma_1}{\Gamma_1, \Gamma_1' \Rightarrow^{[\mathsf{i}]} \psi_1}\ \mathbf{React}$$

$$\begin{array}{c} \dfrac{\Gamma_1, \Gamma_1' \Rightarrow^{[\mathsf{i}]} \Delta \qquad \Gamma_2, \Gamma_2' \Rightarrow^{[\mathsf{e}]} \psi_2 \qquad \psi_2 \Rightarrow^{[*]} \neg \bigwedge \Gamma_1}{\Gamma_1, \Gamma_1' \Rightarrow^{[\mathsf{e}]} \psi_1}\ \mathbf{Retro}^b \\ \vdots \\ \dfrac{\Gamma_2, \Gamma_2' \Rightarrow^{[\mathsf{e}]} \Delta \qquad \Gamma_3 \Rightarrow^{[\mathsf{e}]} \psi_3 \qquad \psi_3 \Rightarrow^{[*]} \neg \bigwedge \Gamma_2}{\Gamma_2, \Gamma_2' \Rightarrow^{[\mathsf{i}]} \psi_2}\ \mathbf{React} \end{array}$$

$$\frac{\Gamma \Rightarrow^{[\mathsf{i}]} \psi \qquad (\forall \Delta \in \mathsf{Def}(\Gamma))\ \Delta \Rightarrow^{[*]} \neg\bigwedge\Gamma \qquad (\forall \Delta \in \mathsf{Def}(\Gamma),\ \exists \Sigma \in \mathsf{Def}(\Delta))\ \Sigma \Rightarrow^{[!]} \neg\bigwedge\Delta}{\Gamma \Rightarrow^{[!]} \psi}\ \mathbf{Final}$$

Figure 7.1: The annotation revision rules **Def**, **React**, **Retro**, and **Final** of AC. In **Final**, $\mathsf{Def}(\Gamma) = \{\Delta \subseteq \mathcal{S} \mid \Delta \vdash \neg\bigwedge\Gamma\}$. Let $\mathcal{S} \subseteq \mathcal{L}$: side-condition ($a$) denotes that $\Gamma_1, \Gamma_1', \Gamma_2 \subseteq \mathcal{S}$; and (b) denotes that $\Gamma_1, \Gamma_1', \Gamma_2, \Gamma_2', \Gamma_3 \subseteq \mathcal{S}$. The rule **Retro** is a *system of rules*, which stipulates that each application of the topmost rule must be followed by an application of a corresponding **React** rule further down in the derivation.

Annotated rules of the base calculus LC. Each sequent $\Gamma \Rightarrow^{[\mathsf{a}]} \Delta$ in an AC derivation is initially introduced through an application of a rule of the underlying base calculus LC. These rules are extended with annotations: both the sequents in the conditions of a LC-rule and the sequent in the conclusion are annotated by [i]. For instance, consider the following *annotated* LC-admissible rule for lhs conjunction introduction, which follows from Definition 7.2:

$$\frac{\Gamma, \varphi, \psi \Rightarrow^{[\mathsf{i}]} \Delta}{\Gamma, \varphi \wedge \psi \Rightarrow^{[\mathsf{i}]} \Delta}$$

Hence, each sequent introduced to a AC-derivation is by default taken as *accepted.*

Attack rules. These are inference rules for changing the annotations of an attacked sequent from [i] to [e]. In this chapter, we consider one such attack rule **Def**, which simulates an *undermining defeat* on a sequent's premises.[1] The rule **Def** is defined as follows:

[1] Further sequent-based attack rules can be imported from, e.g., (Arieli and Straßer, 2019), using annotations similar to **Def**. We leave the consideration of alternative rules for future work.

$$\dfrac{\overbrace{\Gamma_1, \Gamma_1' \Rightarrow^{[\mathsf{i}]} \psi_1}^{\text{attacked}} \qquad \overbrace{\Gamma_2 \Rightarrow^{[\mathsf{i}]} \psi_2}^{\text{attacker}} \qquad \overbrace{\psi_2 \Rightarrow^{[*]} \neg \bigwedge \Gamma_1}^{\text{attack condition}}}{\Gamma_1, \Gamma_1' \Rightarrow^{[\mathsf{e}]} \psi_1}\ \mathbf{Def}^a$$

where (a) stipulates that $\Gamma_1, \Gamma_1', \Gamma_2 \subseteq \mathcal{S}$ for some given set $\mathcal{S} \subseteq \mathcal{L}$. The side-condition (a) restricts **Def** to a set $\mathcal{S}$ of premises, which means that the attacking and the attacked sequent must be $\mathcal{S}$-based. The reason for doing so is to avoid arbitrary attacks. In **Def**, the introduced sequent $\Gamma_2 \Rightarrow^{[\mathsf{i}]} \psi_2$ attacks $\Gamma_1, \Gamma_1' \Rightarrow^{[\mathsf{i}]} \psi_1$, and so changes the latter's status from 'introduced' to 'eliminated'. The derivation status of the attack condition on the right does not matter as long as it is logically valid. Furthermore, the attack condition defines the type of attack. In this case, the attack is on some premises of the attacked sequent in the first condition and, so, the attack undermines. To facilitate readability, in the sequel, we present the attacked sequent in the rule's conditions on the far left, the attacker in the middle, and the attacking condition on the far right.

The **Def** attack rule is accompanied by a variation, in which the attacker $\Gamma_2 \Rightarrow^{[\mathsf{i}]} \psi_2$ is replaced by a finally accepted sequent $\Gamma_2 \Rightarrow^{[!]} \psi_2$. Thus, attacking sequents in **Def** are either accepted or finally accepted. We discuss finally accepted sequents further down below. In what follows, we assume this variation to be implicitly present.

Reactivation rules. These are inference rules that reactivate sequents, changing the annotation of a sequent from eliminated [e] to introduced [i]. In general, each attack rule has a corresponding reactivation rule (cf. footnote 1). The reactivation rule **React**, which corresponds to the attack rule **Def**, is defined as follows:

$$\dfrac{\Gamma_1, \Gamma_1' \Rightarrow^{[\mathsf{e}]} \psi_1 \qquad \Gamma_2 \Rightarrow^{[\mathsf{e}]} \psi_2 \qquad \psi_2 \Rightarrow^{[*]} \neg \bigwedge \Gamma_1}{\Gamma_1, \Gamma_1' \Rightarrow^{[\mathsf{i}]} \psi_1}\ \mathbf{React}$$

The rule indicates that if the sequent in the first condition was previously attacked and one of its attackers is counter-attacked (second condition), then the initially attacked sequent changes its status from 'eliminated' back to 'introduced', i.e., conditionally accepted.[2] The third condition expresses the attacking condition.

Retrospective attack rules. Unlike attack rules that allow only introduced attackers, retrospective attack rules allow for eliminated attackers, provided that the attacker in question can be reactivated.[3] These rules may thus be viewed as pairs of attack and reactivation rules and are similar to systems of rules in sequent-style proof systems (Negri, 2014). The retrospective attack rule **Retro** is defined accordingly:

[2] One may stipulate that the attacker of the sequent must be the sequent through which the attacked sequent was initially eliminated, but this is not necessary for the present purpose.

[3] Intuitively, retrospective attacks deal with attack cycles. We discuss this in Section 7.2, page 269.

$$\frac{\Gamma_1, \Gamma_1' \Rightarrow^{[\mathsf{i}]} \psi_1 \qquad \Gamma_2, \Gamma_2' \Rightarrow^{[\mathsf{e}]} \psi_2 \qquad \psi_2 \Rightarrow^{[*]} \neg \bigwedge \Gamma_1}{\Gamma_1, \Gamma_1' \Rightarrow^{[\mathsf{e}]} \psi_1}\ \textbf{Retro}^{b}$$

(attack rule with the eliminated attacker)

$$\vdots$$

$$\frac{\Gamma_2, \Gamma_2' \Rightarrow^{[\mathsf{e}]} \psi_2 \qquad \Gamma_3 \Rightarrow^{[\mathsf{e}]} \psi_3 \qquad \psi_3 \Rightarrow^{[*]} \neg \bigwedge \Gamma_2}{\Gamma_2, \Gamma_2' \Rightarrow^{[\mathsf{i}]} \psi_2}\ \textbf{React}$$

(the eliminated attacker is reactivated)

where (b) stipulates that $\Gamma_1, \Gamma_1', \Gamma_2, \Gamma_2', \Gamma_3 \subseteq \mathcal{S}$ for some given set $\mathcal{S} \subseteq \mathcal{L}$. Note that the rules above need not be consecutive in the derivation, but the reactivation of the attacker must be part of the, what we call, revision process following the attack rule. A retrospective attack rule is only applicable in case the revision process leads to a reactivation. However, the second **React** rule may be applied irrespective of whether **Retro** has been applied previously (cf. systems of rules (Negri, 2014)). We formally introduce this revision procedure in Definition 7.5. See also the examples in Section 7.2.

Final acceptability rules. These are rules for assuring final inferences of sequents. A finally accepted sequent $s[!]$ corresponds to a skeptical inference in the proof system (see Definition 7.9 below). For now, we consider one such rule: **Final**. Below, we also present several alternative rules, which are all AC-admissible in the light of **Final**. The rule **Final** depends on the attack rule **Def** (which determines the attack condition) and the sequent whose final acceptability is verified. In our case, the attack condition expresses an undermining attack (see **Def** above). Furthermore, **Final** is relative to a *finite* set $\mathcal{S} \subseteq \mathcal{L}$ of premises.[4] Let $\mathsf{Def}(\Gamma) = \{\Delta \subseteq \mathcal{S} \mid \Delta \vdash \neg \bigwedge \Gamma\}$ be the set of all sets of formulae from $\mathcal{S}$ that contradict with the set Γ. The rule **Final** is defined as follows:

$$\frac{\Gamma \Rightarrow^{[\mathsf{i}]} \psi \qquad (\forall \Delta \in \mathsf{Def}(\Gamma))\ \Delta \Rightarrow^{[*]} \neg \bigwedge \Gamma \qquad (\forall \Delta \in \mathsf{Def}(\Gamma),\ \exists \Sigma \in \mathsf{Def}(\Delta))\ \Sigma \Rightarrow^{[!]} \neg \bigwedge \Delta}{\Gamma \Rightarrow^{[!]} \psi}\ \textbf{Final}$$

The **Final** rule signifies that $\Gamma \Rightarrow \psi$ is finally accepted if it is conditionally accepted (the first condition of the rule), all of its attackers are produced in the derivation (this is the second condition of the rule), and each such attacker is counter-attacked by a finally accepted sequent (the last condition of the rule).[5] Notice that the second and third conditions of the rule represent sets of sequents. If the lhs of a sequent $\Delta \Rightarrow^{[*]} \Gamma$ is empty, there are no attackers identified by $\mathsf{Def}(\Delta)$.

[4] Finiteness is required to ensure that the rule does not contain infinitely many conditions.

[5] One can alternatively verify that all attackers are produced in a derivation by employing a refutation calculus $\mathcal{R}$ for which $\Gamma \not\Rightarrow \psi$ is $\mathcal{R}$-provable iff $\Gamma \not\vdash \psi$. See (Bonatti and Olivetti, 2002; Pkhakadze and Tompits, 2020). One may additionally impose a relevance condition on $\mathsf{Def}(\Gamma)$. Namely, by monotonicity on the base logic L it suffices to refer in **Final** to the set $\mathsf{MinAtt}(\Gamma) = \{\Delta \in \mathsf{Def}(\Gamma) \mid (\forall \Delta' \subsetneq \Delta)\, \Delta' \notin \mathsf{Def}(\Gamma)\}$. We will not pursue these options here.

Admissible rules. We highlight that certain additional rules are admissible in light of the above rules. For instance, sequents that cannot be **Def**-attacked by any $\mathcal{S}$-based sequent, either since their left-hand side is empty (these are tautological sequents) or because $\mathcal{S} \notin \mathsf{Def}(\Gamma)$, are finally accepted too. Thus, we also have the following:

$$\frac{\Rightarrow^{[\mathsf{i}]} \psi}{\Rightarrow^{[!]} \psi} \qquad\qquad \frac{\Gamma \Rightarrow^{[\mathsf{i}]} \psi \qquad \mathcal{S} \notin \mathsf{Def}(\Gamma)}{\Gamma \Rightarrow^{[!]} \psi}$$

Further admissible final derivability rules may be expressed. For instance, since we only consider the premise-attack rule **Def**, we can have rules which express that if a sequent is finally derived, so is any derivable sequent with weaker support:

$$\frac{\Gamma, \Gamma' \Rightarrow^{[!]} \varphi \qquad \Gamma \Rightarrow^{[\mathsf{i}]} \psi}{\Gamma \Rightarrow^{[!]} \psi} \qquad \frac{\Gamma_1 \Rightarrow^{[!]} \varphi \qquad \Gamma_1 \Rightarrow^{[\mathsf{i}]} \bigwedge \Gamma_2 \qquad \Gamma_2 \Rightarrow^{[\mathsf{i}]} \psi}{\Gamma_2 \Rightarrow^{[!]} \psi} \text{ with } \Gamma_1, \Gamma_2 \subseteq \mathcal{S}$$

We construct annotated sequent calculi from the above collection of rules.

Definition 7.5 (Annotated Calculi AC)**.** *Let* L *be a base logic with language* $\mathcal{L}$ *and let* LC *be its corresponding sound and complete sequent calculus. An* Annotated Calculus, *referred to as* AC, *consists of the following components:*

- ***The initial and inference rules of*** LC, *where the sequents in the conditions and conclusion of a* LC*-rule are annotated by* $[\mathsf{i}]$*;*
- ***The annotation revision rules Def****,* ***React****,* ***Retro****, and* ***Final*** *from Figure 7.1.*

For any $\mathcal{S} \subseteq \mathcal{L}$, *an* $\mathcal{S}$*-based* derivation $\mathcal{D}$ *of an annotated calculus* AC *is a finite sequence of tuples* $\mathcal{D} = \langle T_1, \ldots, T_n \rangle$ *where the index* $1 \leq i \leq n$ *determines the tuple's order in the derivation* $\mathcal{D}$*. Each* T_i *contains the derived annotated sequent (the tuple's sequent), the derivation rule that is applied (the tuple's rule), and the indexes of the tuples whose sequents serve as the conditions of the tuple's rule. Applications of the rules* ***Def****,* ***Retro****, and* ***Final*** *in* $\mathcal{D}$ *are* $\mathcal{S}$*-dependent.*

We impose the following restrictions of $\mathcal{S}$*-based derivations:*

- *During the construction of a derivation* $\mathcal{D}$*, after each extension with a tuple* T_i *containing an attack rule (****Def****) or a retrospective attack rule (****Retro****), an* annotation revision process *is initiated, and the derivation is extended with new attack or reactivation rules for updating the sequent annotations when necessary.*
- *Reactivation rules are applied only during a revision process.*

Definition 7.7 describes the revision process.

Since the status of sequents may change several times throughout an AC-derivation, we are interested in a sequent's most recent status. In what follows, we make frequent use of the following definition.

Definition 7.6 (Most Recent Status)**.** *Let $\mathcal{D}$ be an* AC*-derivation and let s be an* AC*-sequent . If s*[a1] *is derived in $\mathcal{D}$ and at no later stage in the derivation s*[a2] *is derived, we say that* the most recent status *of s in $\mathcal{D}$ is* [a1]*.*

Definition 7.7 (The Annotation Revision Process)**.** *Let $p(s)$ denote the sequent s*[a] *stripped from its annotation* [a] *and let* RevSeq *consist of the sequents whose annotations are revised during the* revision process*: During the construction of a derivation, when a tuple T_i with a (retrospective) attack on a sequent s is introduced at derivation step i, we let* $\mathsf{RevSeq} = \{p(s)\}$ *and the derivation sequence is traversed backward, starting from the tuple directly preceding T_i. Proceed as follows:*

- *If during the traversal a tuple T_j $(j < i)$ containing a (retrospective) attack is encountered, with attacker s_1 and attacked s_2, such that $p(s_1) \in$* RevSeq *while $p(s_2) \notin$* RevSeq[6]*, then:*
  - *If the most recent status of s_1 is* [e] *(i.e., $p(s_1)$ was (counter-)attacked during the revision), then T_j is revised by means of extending the derivation with a new tuple T_k containing a reactivation rule **React** for $p(s_2)$. As a consequence, the most recent status of $p(s_2)$ in the derivation becomes* [i]*, and $p(s_2)$ is added to* RevSeq*.*
  - *If the most recent status of s_1 is* [i] *(i.e., $p(s_1)$ was reactivated during the revision), and the most recent status of the attacked sequent $p(s_2)$ is also* [i]*, then T_j is revised by means of extending the derivation with a new tuple T_k containing a re-application of the original attack rule **Def** on $p(s_2)$. As a consequence, the most recent status of $p(s_2)$ in the derivation becomes* [e]*, and $p(s_2)$ is added to* RevSeq*.*
- *If during the traversal a reactivating tuple T_j is encountered, where the condition of the reactivation is s_1 and the reactivated sequent is s_2, such that $p(s_1) \in$* RevSeq *and $p(s_2) \notin$* RevSeq*, then:*
  - *If the most recent status of s_1 is* [i] *(i.e., $p(s_1)$ has been reactivated during the traversal), then T_j is revised by means of extending the derivation with a new tuple T_k containing a re-application of the original attack rule on $p(s_2)$. As a consequence, the most recent status of $p(s_2)$ becomes* [e] *and $p(s_2)$ is added to* RevSeq*.*

[6] That is, the status of the attacker has been modified, but the status of the attacked sequent has not been modified yet.

It can be the case that during a revision process initiated by a tuple T, one must traverse the derivation sequence prior to T several times. Nevertheless, this process is finite.

Proposition 7.1. *Let $\mathcal{D}$ be a derivation such that at the last derivation step of $\mathcal{D}$, a tuple T is introduced, initiating a revision process: The revision process is finite.*

Proof. First, observe that $\mathcal{D}$ is finite by Definition 7.5 and thus the derivation sequence preceding T, denoted by $\mathcal{D}'$, is finite too. Consequently $\mathcal{D}'$ contains finitely many sequents $s[\mathsf{a}]$. Observe that by Definition 7.7, RevSeq only increases during the revision process and can only contain sequents occurring in $\mathcal{D}$. Then: Either, at some point during the revision, none of the sequents $p(s_1) \in \mathsf{RevSeq}$ triggers a revision in $\mathcal{D}'$ such that s_1 (retrospectively-)attacks/reactivates s_2 and $p(s_2) \notin \mathsf{RevSeq}$; thus, the revision halts. Or, at some point, all sequents in finite $\mathcal{D}'$ are in RevSeq; thus, the revision halts too. The revision is finite. QED

In what follows, we introduce the notion of a *coherent* derivation. Coherence ensures that problematic odd-cycles of attacks are excluded in the derivation process (cf. coherence in Dung's (1995) argumentation frameworks). In Section 7.2, we discuss some examples with odd-cycles in detail.

Definition 7.8. *Let $s[\mathsf{i}]$ or $s[!]$ be an attacking sequent of a derivation tuple T. Then T is* coherent, *if at the end of the revision process following the introduction of T, the most recent status of s is $[\mathsf{i}]$ or $[!]$.*

A derivation is coherent *if all its tuples are coherent, and there is no tuple such that the most recent statuses of both its attacking and attacked sequents are in $\{[!], [\mathsf{i}]\}$ (where attacks are defined by applications of **Def**, **React**, and **Final**).*

In the above definition, we require the most recent status of an attacking sequent in a coherent tuple to be either $[\mathsf{i}]$ or $[!]$. The reason is that the attacking sequent in the tuple initiating the revision process can be either $s[\mathsf{i}]$ or $s[!]$ (cf. the variations of **Def** on page 260). We discuss this in detail in Remark 7.1 on page 270.

Last, we define the inference relation induced by AC.

Definition 7.9 (Skeptic and credulous inference AC)**.** *Let $\mathcal{S} \subseteq \mathcal{L}$ and let AC be an annotated calculus. We define skeptic inference in AC ($\mid\!\sim^{s}_{\mathsf{AC}}$), respectively credulous inference in AC ($\mid\!\sim^{c}_{\mathsf{AC}}$), accordingly:*

- *$\mathcal{S} \mid\!\sim^{s}_{\mathsf{AC}} \psi$ if there is an $\mathcal{S}$-based AC-derivation $\mathcal{D}$ and $\Gamma \Rightarrow^{[!]} \psi$ is derived in $\mathcal{D}$, with $\Gamma \subseteq \mathcal{S}$.*
- *$\mathcal{S} \mid\!\sim^{c}_{\mathsf{AC}} \psi$ if there is a $\mathcal{S}$-based AC-derivation $\mathcal{D}$ and $\Gamma \Rightarrow^{[i]} \psi$ is derived in $\mathcal{D}$, with $\Gamma \subseteq \mathcal{S}$ and $[\mathsf{i}]$ is the sequent's most recent status.*

Henceforth, we omit the subscript AC.

In other words, skeptic inference corresponds to the derivability of finally accepted sequents and credulous inference corresponds to derivability of accepted sequents. This suffices for now.

7.2 Examples

Below we provide some illustrations and discussions of derivations in Annotated Calculi AC. In all of the examples, we use the proof system of Definition 7.5. To preserve the readability of the examples, we only represent the sequents of tuples occurring in a derivation (complete tuples may be easily reconstructed). What is more, AC-derivations may be represented in various ways. Instead of writing those derivations as lists of tuples, we can often, without loss of generality, represent them as tree-like structures. This is illustrated in Example 7.1 (we come back to this in Section 7.6, page 291).

Example 7.1. *Let* LC *be an adequate calculus for classical logic and let* $s_1 = p \Rightarrow p$ *and* $s_2 = \neg p \Rightarrow \neg p$*. The* AC*-derivation,*

$$\begin{aligned}
T_1 &= (s_1[\mathsf{i}], \boldsymbol{Ax}, \emptyset)\\
T_2 &= (s_2[\mathsf{i}], \boldsymbol{Ax}, \emptyset)\\
T_3 &= (s_2[\mathsf{i}], \boldsymbol{Ax}, \emptyset)\\
T_4 &= (s_1[\mathsf{e}], \boldsymbol{Def}, \langle 1, 2, 3\rangle)\\
T_5 &= (s_2[\mathsf{i}], \boldsymbol{Ax}, \emptyset)\\
T_6 &= (s_1[\mathsf{i}], \boldsymbol{Ax}, \emptyset)\\
T_7 &= (s_1[\mathsf{i}], \boldsymbol{Ax}, \emptyset)\\
T_8 &= (s_2[\mathsf{e}], \boldsymbol{Def}, \langle 5, 6, 7\rangle)\\
T_9 &= (s_2[\mathsf{i}], \boldsymbol{Ax}, \emptyset)\\
T_{10} &= (s_1[\mathsf{i}], \boldsymbol{React}, \langle 4, 8, 9\rangle)
\end{aligned}$$

can be represented by the tree-like derivation,

$$\dfrac{\dfrac{\dfrac{}{s_1[\mathsf{i}]}\,\boldsymbol{Ax} \quad \dfrac{}{s_2[\mathsf{i}]}\,\boldsymbol{Ax} \quad \dfrac{}{s_2[\mathsf{i}]}\,\boldsymbol{Ax}}{s_1[\mathsf{e}]}\,\boldsymbol{Def} \quad \dfrac{\dfrac{}{s_2[\mathsf{i}]}\,\boldsymbol{Ax} \quad \dfrac{}{s_1[\mathsf{i}]}\,\boldsymbol{Ax} \quad \dfrac{}{s_1[\mathsf{i}]}\,\boldsymbol{Ax}}{s_2[\mathsf{e}]}\,\boldsymbol{Def} \quad \dfrac{}{s_2[\mathsf{i}]}\,\boldsymbol{Ax}}{s_1[\mathsf{i}]}\,\boldsymbol{React}$$

where the root of the proof is the derivation's conclusion, and the leaves are the initial sequents.

In the above example, the rule **Ax** signifies that the introduced sequent is an initial sequent of LC. In what follows, we often omit the inferences from the underlying calculus

LC and focus instead on the annotation rules. Furthermore, we point out that different orderings of the tuples may give rise to identical tree-like representation.[7] However, the notion of coherence ensures that problematic deviations in orderings of a derivation are excluded.[8] In the remainder of this chapter, we only consider examples that have a tree-like structure. In Section 7.6, we discuss some future work concerning AC and derivation trees.

Example 7.2. *Consider the set of assumptions $\mathcal{S} = \{p, \neg p, q\}$ and let* L *be classical logic. To see that $q \Rightarrow q$ is finally accepted from $\mathcal{S}$ (i.e., there is a derivation of $q \Rightarrow^{[\mathsf{!}]} q$), note that $q \Rightarrow^{[\mathsf{i}]} q$ and $\Rightarrow^{[\mathsf{i}]} p \vee \neg p$ are derivable (e.g., using Gentzen's sequent calculus* LK*). Moreover, by the final acceptability rule* ***Final****, $\Rightarrow p \vee \neg p$ is finally accepted (since its left-hand side is empty and thus cannot be attacked). Now, the $\mathcal{S}$-based attackers of $q \Rightarrow q$ are $p, \neg p \Rightarrow \neg q$ and $p, \neg p, q \Rightarrow \neg q$ and so* $\mathsf{Att}(q) = \{\{p, \neg p\}, \{p, \neg p, q\}\}$. *These attackers are also derivable, but $\Rightarrow^{[\mathsf{!}]} p \vee \neg p$ attacks both of them. For instance, we have,*

$$\dfrac{p, \neg p \Rightarrow^{[\mathsf{i}]} \neg q \qquad \Rightarrow^{[\mathsf{!}]} p \vee \neg p \qquad p \vee \neg p \Rightarrow^{[\mathsf{i}]} \neg(p \wedge \neg p)}{p, \neg p \Rightarrow^{[\mathsf{e}]} \neg q}\ \boldsymbol{\mathit{Def}}$$

Consequently, by the final acceptability rule ***Final*** *we derive $q \Rightarrow^{[\mathsf{!}]} q$ and so $\mathcal{S} \mathrel{|\!\sim}^s q$.*

The situation regarding the other formulas in $\mathcal{S}$ is different, and we have that $\mathcal{S} \mathrel{|\!\not\sim}^s p$ and $\mathcal{S} \mathrel{|\!\not\sim}^s \neg p$. Indeed, while both $p \Rightarrow^{[\mathsf{i}]} p$ and $\neg p \Rightarrow^{[\mathsf{i}]} \neg p$ are derivable, none of them is finally accepted. This may be explained by the fact that each attacks the other, causing an iterated revision of their statuses. Indeed, after

$$\dfrac{\neg p \Rightarrow^{[\mathsf{i}]} \neg p \qquad p \Rightarrow^{[\mathsf{i}]} p \qquad p \Rightarrow^{[\mathsf{i}]} \neg\neg p}{\neg p \Rightarrow^{[\mathsf{e}]} \neg p}\ \boldsymbol{\mathit{Def}}$$

$p \Rightarrow p$ is accepted and $\neg p \Rightarrow \neg p$ is eliminated. However, by subsequently extending the derivation with the retrospective attack

$$\begin{array}{c}
\dfrac{p \Rightarrow^{[\mathsf{i}]} p \qquad \neg p \Rightarrow^{[\mathsf{e}]} \neg p \qquad \neg p \Rightarrow^{[\mathsf{i}]} \neg p}{p \Rightarrow^{[\mathsf{e}]} p}\ \boldsymbol{\mathit{Retro}} \\
\vdots \\
\dfrac{\neg p \Rightarrow^{[\mathsf{e}]} \neg p \qquad p \Rightarrow^{[\mathsf{e}]} p \qquad p \Rightarrow^{[\mathsf{i}]} \neg\neg p}{\neg p \Rightarrow^{[\mathsf{i}]} \neg p}\ \boldsymbol{\mathit{React}}
\end{array}$$

[7] One could enhance the tree-like derivation with numbers assigned to sequents, indicating the order of construction of the proof.

[8] This means that AC-derivability is not yet invariant under the order of proof's construction. We address this when we discuss future work in Section 7.6.

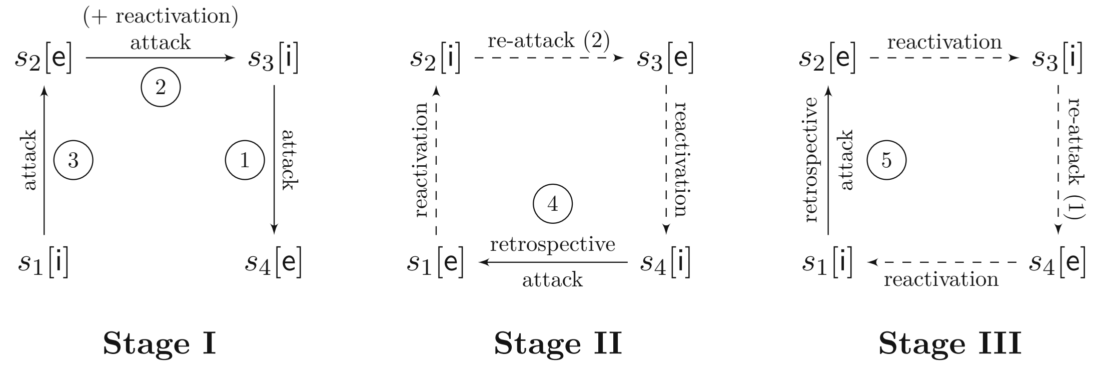


Figure 7.2: Derivations for an attack cycle of length four (Example 7.3). The numbers represent the order of the derivation steps. Strict arrows are applications of attack rules, and dashed arrows denote rules applied in the corresponding annotation revision process. The terms attack, retrospective attack, and reactivation correspond to applications of the rules **Def**, **Retro**, and **React**, respectively.

the situation is reversed, and now $\neg p \Rightarrow \neg p$ is accepted while $p \Rightarrow p$ is eliminated. Another application of the retrospective attack rule, this time with $p \Rightarrow^{[\mathsf{e}]} p$ retrospectively attacking $\neg p \Rightarrow^{[\mathsf{i}]} \neg p$, again reverses their statuses, and so forth.

Example 7.3. *The previous example demonstrates a cyclic attack of size two where $p \Rightarrow p$ and $\neg p \Rightarrow \neg p$ reciprocally attack each other. Figure 7.2 shows a cycle of size four. The left-hand side of the figure, Stage I, represents a 3-chain of attacks. In the middle of the figure (Stage II), the chain turns into a 4-cycle. Note that in order to close the cycle, a retrospective attack is needed. After the revision process (denoted by the dashed arrows), s_2 and s_4 are accepted, while s_1 and s_3 are eliminated. These statuses may be reversed by an application of another retrospective attack, as shown in Stage III of the figure. As a result, s_1 and s_3 are accepted while s_2 and s_4 are eliminated. The last two steps may be repeated interchangeably, revising each time the statuses of the sequents involved in the cycle of attacks. It follows that each of the four sequents is exposed to repeated attacks, so none is finally accepted. Still, we find that the statuses of s_1 and s_3 are inevitably linked, and the same holds for s_2 and s_4. That is, these pairs are either jointly accepted or jointly eliminated.*

A similar analysis holds for any even-length cycle of attacks. Thus derivation tuples in even-length cycles are coherent *(Definition 7.8). In contrast, derivations with an odd-length cyclic attack are incoherent. See, e.g., derivation step 3 in Figure 7.3 Stage II, for a 3-cycle. Indeed, for 'closing' such a cycle with an attack rule, the attacker must be eliminated at the end of the revision process (this may be verified by induction on the length of the cycle). Moreover, if a retrospective attack is initiated in such cases, the attacker cannot be reactivated.*

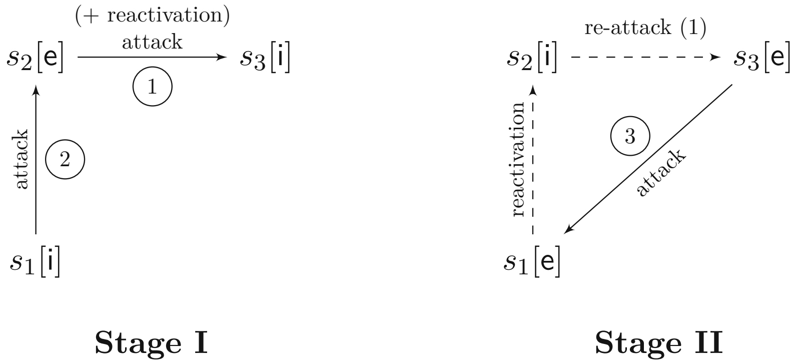


Figure 7.3: Derivations for an attack cycle of length three (Example 7.3). The numbers represent the order of the derivation steps. Strict arrows are applications of attack rules, and dashed arrows denote rules applied in the corresponding annotation revision process. The terms attack and reactivation correspond to applications of the rules **Def** and **React**, respectively.

Example 7.4. *Consider a logic whose negation $\neg$ does not respect double-negation introduction (i.e., $p \not\vdash \neg\neg p$), and suppose again that **Def** is the only attack rule. Let $\mathcal{S} = \{p, \neg p, \neg\neg p, \neg\neg\neg p\}$. We denote by $\neg_i p$ the formula in which p is preceded by i-many negations (in particular, $\neg_0\, p$ is p) and by $s_i[\mathsf{a}]$ the annotated sequent $\neg_i p \Rightarrow^{[\mathsf{a}]} \neg_i p$ for $\mathsf{a} \in \{\mathsf{i}, \mathsf{e}, !\}$. By reflexivity, $s_i[\mathsf{i}]$ is derivable for every $0 \leq i \leq 3$. Now, consider the following coherent derivation:*

$$\dfrac{s_0[\mathsf{i}] \qquad \dfrac{\dfrac{s_1[\mathsf{i}] \quad s_2[\mathsf{i}] \quad s_2[\mathsf{i}]}{s_1[\mathsf{e}]}\,\textbf{\textit{Def}} \qquad \dfrac{s_2[\mathsf{i}] \quad s_3[\mathsf{i}] \quad s_3[\mathsf{i}]}{s_2[\mathsf{e}]}\,\textbf{\textit{Def}} \qquad s_2[\mathsf{e}]}{s_1[\mathsf{i}]}\,\textbf{\textit{React}} \qquad s_1[\mathsf{i}]}{s_0[\mathsf{e}]}\,\textbf{\textit{Def}}$$

*At this point, the derived sequents and their most updated statuses are $s_0[\mathsf{e}]$, $s_1[\mathsf{i}]$, $s_2[\mathsf{e}]$ and $s_3[\mathsf{i}]$. (We discuss the order of tuples in a derivation in more detail in Section 7.6.) Moreover, this is a kind of a 'stable state', in which all the sequents that are conditionally accepted (and only those) can, in fact, be finally accepted (note that all their $\mathcal{S}$-based attackers are derived). Indeed, since $\mathcal{S} \notin \mathsf{Att}(\neg_3 p)$, we have that $\mathsf{Att}(\neg_3 p) = \emptyset$ (i.e., s_3 cannot be attacked by any $\mathcal{S}$-based sequent), and so by the final acceptability rule **Final** we can further extend the derivation and derive $s_3[!]$. In turn, $s_3[!]$ attacks s_2, the single attacker of s_1, thus $s_1[!]$ is derived too. Finally, $s_1[!]$ attacks s_0, and so the latter cannot be finally accepted. It follows, then, that $\mathcal{S} \mathrel{|\!\sim}^{s} \neg p$ and $\mathcal{S} \mathrel{|\!\sim}^{s} \neg\neg\neg p$.*

Now, suppose that $\neg_4 p$ is added to the set of assertions, resulting in the set $\mathcal{S}' = \{p, \neg p, \neg\neg p, \neg\neg\neg p, \neg\neg\neg\neg p\}$. Then $s_4[\mathsf{i}]$ is derived, and so the previous derivation may be extended as follows:

$$\dfrac{s_0[\mathsf{e}] \quad \dfrac{s_1[\mathsf{i}] \quad \dfrac{s_2[\mathsf{e}] \quad \dfrac{s_3[\mathsf{i}] \quad s_4[\mathsf{i}] \quad s_4[\mathsf{i}]}{s_3[\mathsf{e}]}\,\textbf{\textit{Def}} \quad s_3[\mathsf{e}]}{s_2[\mathsf{i}]}\,\textbf{\textit{React}} \quad s_2[\mathsf{i}]}{s_1[\mathsf{e}]}\,\textbf{\textit{Def}} \quad s_1[\mathsf{e}]}{s_0[\mathsf{i}]}\,\textbf{\textit{React}}$$

In the resulting coherent derivation, the derived sequents and their most updated statuses are now $s_0[\mathsf{i}]$, $s_1[\mathsf{e}]$, $s_2[\mathsf{i}]$, $s_3[\mathsf{e}]$ *and* $s_4[\mathsf{i}]$. *Again, for similar reasons as before, all the sequents that are conditionally accepted (and only them) are also finally accepted, so this time we conclude that* $\mathcal{S}' \mid\!\sim^s p$ *and* $\mathcal{S}' \mid\!\sim^s \neg\neg p$ *and* $\mathcal{S}' \mid\!\sim^s \neg\neg\neg\neg p$. *An alternative representation of this derivation is presented in Figure 7.4.*

7.3 Basic Properties of Annotated Calculi

Next, we consider some basic properties of the consequence relation $\mid\!\sim^s$ induced by annotated calculi AC, thus addressing Objective 1. We start with some observations about derivations in annotated calculi.

Proposition 7.2. *Let* $\mathcal{D}$ *be an* AC*-derivation. For a fixed set of assumptions, a finally accepted sequent cannot be eliminated.*

Proof. This result holds because a finally accepted sequent cannot be attacked (cf. **Def** and **Retro**). QED

Remark 7.1. *Although the sequent* $s[!]$ *itself cannot be eliminated (Proposition 7.2), it may be the case that the most recent status of* s *changes throughout the derivation. Recall that* AC *contains two types of* ***Def*** *rules: one where the status of the attacking sequent* s *is* $[\mathsf{i}]$ *and one where its status is* $[!]$. *The definition of coherence ensures that after a revision process, the status of the attacking sequent* s *is* $[\mathsf{i}]$ *or* $[!]$. *However, it may be the case that the attacking sequent's status changes from* $[!]$ *to* $[\mathsf{i}]$ *(for instance, during the revision process). The following coherent derivation illustrates this. For the sake of readability, we omit the attacking condition of* ***Def*** *and* ***React*** *rules, which in all cases is the same as the attacking argument in the rule's application.*

$$\dfrac{\dfrac{s_1[\mathsf{i}] \quad s_2[\mathsf{i}]}{s_1[\mathsf{e}]}\,\textbf{\textit{Def}} \quad \dfrac{s_2[\mathsf{i}] \quad \dfrac{\dfrac{\dfrac{s_1[\mathsf{i}] \quad s_2[\mathsf{i}]}{s_1[\mathsf{e}]}\,\textbf{\textit{Def}} \quad \dfrac{s_2[\mathsf{i}] \quad s_1[\mathsf{i}]}{s_2[\mathsf{e}]}\,\textbf{\textit{Def}}}{s_1[\mathsf{i}]}\,\textbf{\textit{React}} \quad s_2[\mathsf{i}] \quad \dfrac{s_3[\mathsf{i}] \quad \emptyset \quad \emptyset}{s_3[!]}\,\textbf{\textit{Final}}}{s_1[!]}\,\textbf{\textit{Final}}}{s_2[\mathsf{e}]}\,\textbf{\textit{Def}}}{s_1[\mathsf{i}]}\,\textbf{\textit{React}}$$

We can even prove a stronger result. Namely, Proposition 7.3 expresses that no derivation $\mathcal{D}$ can be coherently extended (by further derivation steps and revisions), based on the

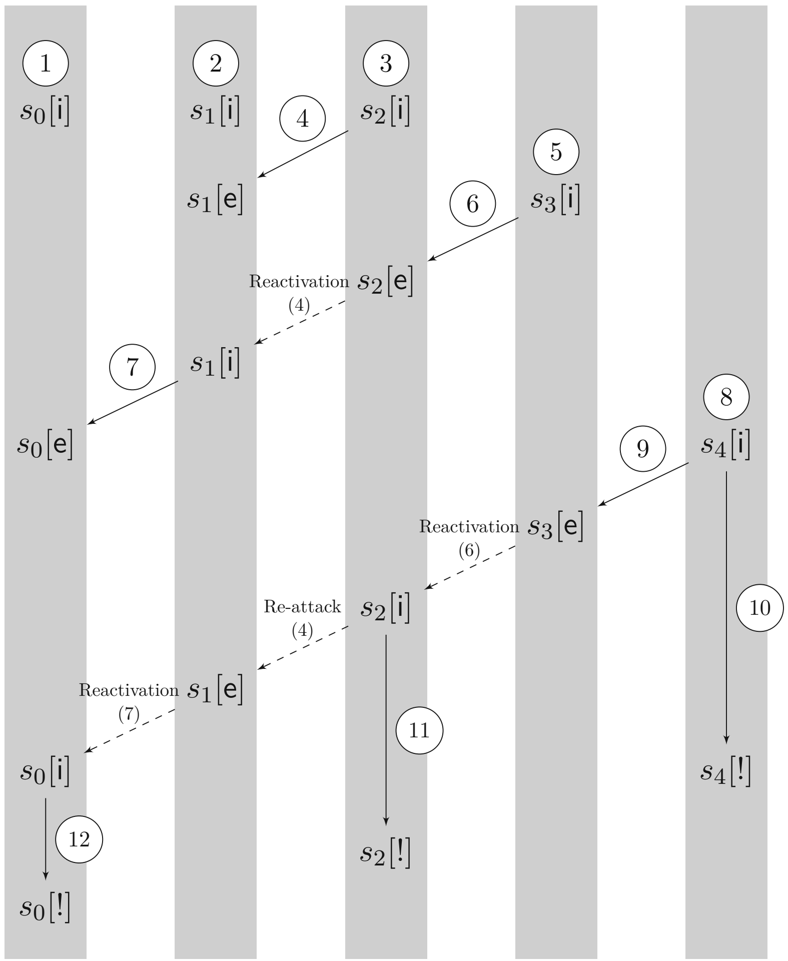


Figure 7.4: The derivation of Example 7.4 for $\mathcal{S}'$ (progressing along the vertical axis according to the circled numbers, which represent derivation steps). The solid arrows are applications of attack rules, and the dashed arrows are the rules applied in the corresponding annotation revision process. The gray rectangles highlight the status changing of each sequent.

same set of assumptions, such that the *most recent status* of a sequent $s[!] \in \mathcal{D}$ is [e]. In other words, once a sequent s has been derived as finally accepted somewhere in the derivation, it remains (finally) accepted in any coherent extension. This resembles the notion of final derivability in adaptive logics (Batens, 2007; Straßer, 2014) and in (Arieli and Straßer, 2019). As a corollary, we thus know that any attacker of a finally accepted sequent s is *permanently eliminated*. In Section 7.5, we extend the framework with the annotation [$\perp$] expressing permanently eliminated sequents.

Proposition 7.3. *Let $\mathcal{D}$ be a coherent derivation with a fixed set of assumptions and let $s[!] \in \mathcal{D}$. The most recent status of s is either* [!] *or* [i].

Proof. Let $\langle s_1, \ldots, s_n \rangle$ be the ordered set of all [!]-sequents in the order of their occurrence in $\mathcal{D}$. We prove by induction of n that the most recent status of s_i $(1 \leq i \leq n)$ is either [!] or [i]. *Base case.* Since s_1 is the first derived s[!] sequent we know it has no attackers. Hence, there is no possible attack rule applicable to s and, so, the most recent status of s in $\mathcal{D}$ is [!]. *Inductive step.* Suppose s_{i+1} was derived by an application of **Final** using s_{j_1}[!], $\ldots, s_{j_m}$[!]. By IH we know that,

$$(\dagger) \text{ the most recent status of } s_{j_k} \ (1 \leq k \leq m) \text{ is [!] or [i].}$$

Suppose towards a contradiction that the most recent status of s_{i+1} is [e] at derivation step l. Hence, **Def** was applied to s_{i+1}[i] (by Proposition 7.2) with attacker r[i] or r[!] either (1) during a derivation step or (2) during a revision process. We consider both cases.

(1) The application of **Def** initiates a revision process. By Definition 7.7, we know that the status of s_{i+1}[e] did not change during the revision process. By coherence of $\mathcal{D}$, we know that the status of r did not change to [e] during the revision process and hence is either r[i] or r[!]. There are two options. If r[i] or r[!] is the most recent status of r then the derivation $\mathcal{D}$ is not coherent, since r is attacked by some s_{j_k} whose most recent status by (†) is [!] or [i]. If r[i] or r[!] is not the most recent status of r, then we know r[e] has been derived further down in the derivation at a step $l' > l$ (after the end of the revision process initiated at l) through an application of **Def**. Again, **Def** is applied (a) either during a derivation step or (b) during a revision process. Consider (a): Then, at the start of the revision process $\{r\} = \mathsf{RevSeq}$ and since $s_{i+1} \notin \mathsf{RevSeq}$ we know **React** has been applied to change the annotation of s_{i+1} to [i] at some step $l' > l$. This contradicts the assumption that the most recent status of s_{i+1} occurs at l. Consider (b): Then, $r \in \mathsf{RevSeq}$. If $s_{i+1} \notin \mathsf{RevSeq}$, we obtain a contradiction similar to item (a). If $s_{i+1} \in \mathsf{RevSeq}$, then this means that the status of s_{i+1} was changed during the revision process at some step $l' > l$, which contradicts the assumption that the most recent status of s_{i+1} occurred at l.

(2) The application of **Def** is part of a revision process. Hence, $r \in \mathsf{RevSeq}$ and $s_{i+1} \in \mathsf{RevSeq}$. By Definition 7.7, their status did not change again during the revision. Again, there are two options. If r[i] or r[!] is the most recent status of r then the derivation $\mathcal{D}$ is not coherent, since r is attacked by some s_{j_k} whose most recent status by (†) is [!] or [i]. If r[i] or r[!] is not the most recent status of r, then we know r[e] has been derived further down in the derivation at a step $l' > l$ (after the end of the revision process) through an application of **Def**. We then obtain a contradiction by the same reasoning as for (a) and (b) of item (1).

Hence, the most recent status of s_{i+1} is [i] or [!]. QED

Corollary 7.1. *Let $\mathcal{D}$ be a coherent derivation with a fixed set of assumptions and let s attack r somewhere in $\mathcal{D}$ such that s[!] $\in \mathcal{D}$. The most recent status of r is* [e].

The following proposition demonstrates that the coherence of the revision process avoids problematic odd-cycles during the process (recall Example 7.3 of Section 7.2.

Proposition 7.4. *Let $\mathcal{D}$ be a coherent derivation (Definition 7.8). At the end of a revision process initiated by a (retrospective-)attack rule, no sequent attacks another sequent during the revision and is eliminated at the same time.*

Proof. Suppose that s_1 attacks s_2. This may happen in one of the following two cases:

1. The attack is part of the derivation step initiating the revision. If this step contains an application of the attack rule **Def**, then the status of s_1 is $[\mathsf{i}]$ or $[!]$, and thus by coherence s_1 cannot be eliminated during the revision. The other possibility is that a retrospective attack rule **Retro** is applied, in which case s_1 must be reactivated during the revision, turning its status back to $[\mathsf{i}]$. Since $s_1 \in \mathsf{RevSeq}$ after the reactivation, its status cannot be modified again in the revision process.

2. The attack is reinstated as part of the revision process. This may happen only if s_1 was reactivated earlier in the revision process and, so, its status was changed to $[\mathsf{i}]$. Since $s_1 \in \mathsf{RevSeq}$, its status cannot be changed again during the revision.

In both cases, s_1 is not eliminated. QED

Next, we demonstrate that in the case of a *consistent* set of assertions $\mathcal{S} \subseteq \mathcal{L}$ the consequence relation $\mathrel{|\!\sim}^{s}$ coincides with the monotonic consequence relation $\vdash$ of the underlying base logic. This is expressed by Proposition 7.5. Thereafter, we show that in case of a set of inconsistent assertions $\mathcal{S}$, the consequence relation $\mathrel{|\!\sim}^{s}$ behaves like a nonmonotonic inference relation, satisfying *paraconsistency*. With a paraconsistent consequence relation, we mean that from an inconsistent set $\mathcal{S} \subseteq \mathcal{L}$, we do not obtain the explosive conclusion set $\mathcal{L}$. This property is demonstrated in Proposition 7.6.

Proposition 7.5. *Let* LC *be a calculus,* $\mathcal{L}$ *its language, and let* AC *be an annotated calculus based on* LC*. If $\mathcal{S} \subseteq \mathcal{L}$ is $\vdash$-consistent (i.e., $\mathcal{S} \nvdash \neg\psi$ for every $\psi \in \mathcal{S}$), then $\mathcal{S} \vdash \psi$ iff $\mathcal{S} \mathrel{|\!\sim}^{s} \psi$.*[9]

Proof. If $\mathcal{S}$ is $\vdash$-consistent, no attack rule is applied. Thus no sequent is eliminated (and so no reactivation or retrospective attack is applied either). It follows that in this case, a derivation consists only of rules of LC and the final acceptability rules. Moreover, for every $\Gamma \subseteq \mathcal{S}$ it holds that $\mathsf{Att}(\Gamma) = \emptyset$, thus any derived sequent is also finally accepted (and, of course, any finally accepted sequent must be derived). It follows that $\mathcal{S} \vdash \psi$ iff $\mathcal{S} \Rightarrow^{[\mathsf{i}]} \psi$ is derived, iff $\mathcal{S} \Rightarrow \psi$ is finally accepted in that derivation, iff $\mathcal{S} \mathrel{|\!\sim}^{s} \psi$. QED

[9] Recall, $\vdash$ is the consequence relation of the base logic and $\mathrel{|\!\sim}^{s}$ is the consequence relation AC given by final acceptance.

Proposition 7.6. *Let* LC *be a base calculus and* AC *an annotated calculus based on* LC*. If* $\vdash$ *is paraconsistent (i.e.,* $p, \neg p \not\vdash q$*) or contrapositive (i.e., if* $\Gamma \vdash \psi$ *then* $\Gamma, \neg\psi \vdash \neg\gamma$ *for every* $\gamma \in \Gamma$*), then* $\mid\!\sim^{s}$ *is paraconsistent .*

Proof. If $\vdash$ is paraconsistent, then $p, \neg p \Rightarrow^{[\mathsf{i}]} q$ is not LC-derivable and fortiori not AC-derivable. Since neither $p \Rightarrow^{[\mathsf{i}]} q$ nor $\neg p \Rightarrow^{[\mathsf{i}]} q$ is LC-derivable (the logic is not trivial), there is no derivable $\{p, \neg p\}$-based sequent whose conclusion is q. Thus, no such sequent is finally derived in AC and so $p, \neg p \not\mid\!\sim^{s} q$. Suppose then that $\vdash$ is not paraconsistent. Then $p, \neg p \Rightarrow^{[\mathsf{i}]}$ is LC-derivable (by LC), and so by contraposition $\Rightarrow \neg(p \wedge \neg p)$ is also LC-derivable. The latter can be derived as finally accepted in AC (by the final acceptability rule **Final**). Moreover, $\{p, \neg p\} \in \mathsf{Att}(q)$. Thus, even if $p, \neg p \Rightarrow^{[\mathsf{i}]} q$ is AC-derived, the last condition in the final acceptability rule is not met. So, $\Gamma \Rightarrow q$ is not finally derived in AC for $\Gamma \subseteq \{p, \neg p\}$, thus $p, \neg p \not\mid\!\sim^{s} q$ in this case as well. QED

Last, we show that the consequence relation of AC is nonmonotonic. We say that a consequence relation is nonmonotonic whenever it does not satisfy the property of monotonicity as given in Definition 7.2.

Proposition 7.7. *If* $\vdash$ *is paraconsistent or contrapositive, the consequence relation* $\mid\!\sim^{s}$ *of* AC *is nonmonotonic.*

Proof. It suffices to consider an example. Consider the set $\mathcal{C} = \{p\} \subseteq \mathcal{L}$. Clearly, $\mathcal{S} \mid\!\sim^{s} p$ since the derivable sequent $p \Rightarrow p$ cannot be attacked relative to $\mathcal{S}$ (the logic is non-trivial and $\mathcal{S}$ is consistent). A proper extension of $\mathcal{S}$ resulting in $\mathcal{S}' = \{p, \neg p\}$ results in $\mathcal{S}' \not\mid\!\sim^{s} p$ by similar reasoning to Proposition 7.6. QED

7.4 Relations to Formal Argumentation

Annotated calculi, particularly the attack rules, are inspired by similar concepts in formal argumentation (Dung, 1995; Baroni et al., 2018; Gabbay et al., 2021). For instance, the attack rule **Def** is similar to undermining attacks defined by (Toulmin, 1958; Pollock, 1987). The reactivation rule **React** is related to the argumentative notion of reinstatement (Baroni et al., 2011). This section shows some relations between the two formalisms, addressing Objective 2: we prove a correspondence between *credulous* and *skeptical* AC-inference and stable and grounded semantics. First, we show under which conditions the derivability of an accepted sequent $s[\mathsf{i}]$ corresponds to membership of a stable extension set of a AC-induced argumentation framework. Second, we show that the derivability of a finally accepted sequent $s[!]$ yields membership of the grounded extension set of a AC-induced argumentation framework.

Definition 7.10 (Argumentation Frameworks induced by AC)**.** *Let* AC *be an annotated sequent calculus, let* $\mathcal{S} \subseteq \mathcal{L}$*, and let* $\mathcal{D}$ *be an* $\mathcal{S}$*-based* AC*-derivation. Then:*

- $\mathsf{Derived}(\mathcal{D})$ *is the set of* $\mathcal{S}$*-based sequents* s *s.t.* $s[\mathsf{i}] \in \mathcal{D}$*;*

- $\mathsf{Accept}(\mathcal{D})$ *is the set of sequents* s *in* $\mathsf{Derived}(\mathcal{D})$ *such that their most updated status is* [i] *or* [!];

- $\mathsf{Final}(\mathcal{D})$ *is the set of sequents s in* $\mathsf{Derived}(\mathcal{D})$ *such that* $s[!] \in \mathcal{D}$;

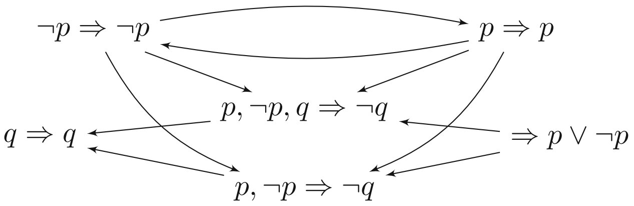
- $\mathsf{Att}(\mathcal{D})$ *is the set of pairs* (s_1, s_2) *such that* s_1 *attacks* s_2 *by an application of either* ***Def***, ***React***, ***Retro***, *or* ***Final*** *in* $\mathcal{D}$, *with* $s_1, s_2 \in \mathsf{Derived}(\mathcal{D})$.

$\mathcal{AF}(\mathcal{D}) = \langle \mathsf{Derived}(\mathcal{D}), \mathsf{Att}(\mathcal{D}) \rangle$ *is called the (sequent-based) argumentation framework induced by* $\mathcal{D}$.

Concerning $\mathsf{Att}(\mathcal{D})$, we point out that the rule **Final** may contain a list of attackers. Namely, the second condition of the rule enumerates all attackers s_i of the sequent s in the first condition, and the third condition of the rule enumerates for each attacker s_i in the second condition a finally derived attacker s_j.

Furthermore, notice that the set $\mathsf{Final}(\mathcal{D})$ not only collects all sequents whose most recent status is [!] but, in fact, contains all sequents whose status is [!] somewhere in the derivation $\mathcal{D}$. Recall Remark 7.1 illustrating that the most recent status of a s[!] may be [i] (even though it can never be eliminated, as shown in Proposition 7.3).

Example 7.5. *The possible attacks among the sequents in Example 7.2 and the argumentation framework induced by the corresponding derivation are the following:*


$\neg p \Rightarrow \neg p$
$p \Rightarrow p$
$p, \neg p, q \Rightarrow \neg q$
$q \Rightarrow q$
$\Rightarrow p \vee \neg p$
$p, \neg p \Rightarrow \neg q$

Definition 7.10 is a logic-based representation of argumentation frameworks, which following Dung (1995) are pairs $\mathcal{AF} = \langle \mathsf{Args}, \mathsf{Att} \rangle$, where Arg is a denumerable set arguments $a, b, c, \ldots$, and Att is a relation on $\mathsf{Arg} \times \mathsf{Arg}$, whose instances are called attacks. Given a framework $\mathcal{AF}$, a key issue in its understanding is what combinations of arguments can *collectively be accepted* in $\mathcal{AF}$. This is determined by various semantic definitions (Baroni et al., 2011) (see Chapter 6 for a discussion of formal argumentation and semantic extensions).

Definition 7.11 (Semantics and Nonmonotonic Inference)**.** *Let* $\mathcal{AF} = \langle \mathsf{Arg}, \mathsf{Att} \rangle$ *be an argumentation framework, and let* $\mathcal{E} \subseteq \mathsf{Arg}$.

- $\mathcal{E}$ attacks *an argument a if there is an argument* $b \in \mathcal{E}$ *that attacks a, i.e.,* $(a, b) \in \mathsf{Att}$. *The set of arguments attacked by* $\mathcal{E}$ *is denoted by* $\mathcal{E}^+$;

- $\mathcal{E}$ defends a *if* $\mathcal{E}$ *attacks every argument that attacks* a*;*
- $\mathcal{E}$ *is* conflict-free *if it does not attack any of its elements, i.e.,* $\mathcal{E}^+ \cap \mathcal{E} = \emptyset$*;*
- $\mathcal{E}$ *is* admissible *if it is conflict-free and defends all of its elements;*
- $\mathcal{E}$ *is* complete *if it is admissible and contains all the arguments it defends;*
- $\mathcal{E}$ *is a* stable extension *of* $\mathcal{AF}$ *if it is conflict-free and* $\mathcal{E} \cup \mathcal{E}^+ = \mathsf{Arg}$*;*
- $\mathcal{E}$ *is a* grounded extension *of* $\mathcal{AF}$ *if it is* $\subseteq$*-minimal among the complete extensions of* $\mathcal{AF}$*.*

Let $\mathtt{Stable}$ *(*$\mathtt{Grounded}$*) be the set of stable (grounded) extensions of* $\mathcal{AF}$*.*[10] *Let* $\mathtt{sem} \in \{\mathtt{Stable}, \mathtt{Grounded}\}$*, we define skeptic (s) and credulous (c) nonmonotonic inference over* $\mathcal{AF}$ *as follows:*

- $\mathcal{AF} \mid\!\sim^{s}_{\mathtt{sem}} \varphi$ *iff for each* $\mathcal{E} \in \mathtt{sem}$*, there is an* $a \in \mathcal{E}$ *concluding* φ*;*
- $\mathcal{AF} \mid\!\sim^{c}_{\mathtt{sem}} \varphi$ *iff there is an* $\mathcal{E} \in \mathtt{sem}$ *s.t. there is an* $a \in \mathcal{E}$ *concluding* φ*.*

We start with credulous inference. The following result shows the close relation between acceptable sequents in AC-derivations $\mathcal{D}$ and credulous inference over stable semantics of argumentation frameworks induced by $\mathcal{D}$.

Proposition 7.8. *Let* $\mathcal{D}$ *be coherent* AC*-derivation:* $\mathsf{Accept}(\mathcal{D})$ *is a stable extension of* $\mathcal{AF}(\mathcal{D})$*.*

Proof. It suffices to show that $\mathsf{Accept}(\mathcal{D})$ is conflict-free in $\mathcal{AF}(\mathcal{D})$ and attacks any eliminated sequent. The former follows from the assumed coherence of $\mathcal{D}$. For the latter, let r be an eliminated sequent in $\mathcal{AF}(\mathcal{D})$. This means that the most recent status of r in $\mathcal{D}$ is $r[\mathsf{e}]$. Consequently, either **Def** or **Retro** was applied at the corresponding derivation step i to derive $r[\mathsf{e}]$. Hence, there is an attacker s of r in the conditions of the rule. In both cases, the rule either (1) initiated a revision process or (2) is part of a revision process.

(1) We first consider the two cases in which a revision process was initiated at i:

Def Then, $s[\mathsf{i}]$ or $s[!]$ is in the conditions of **Def** at step i (recall that we have two versions of **Def**). By coherence of $\mathcal{D}$ this means that the status of s is $[\mathsf{i}]$ or $[!]$ at $j \geq i$ after revision process (cf. Remark 7.1). We prove that the *most recent* status of s is still $[\mathsf{i}]$ or $[!]$ and thus $s \in \mathsf{Accept}(\mathcal{D})$. Suppose towards a contradiction that this is not the case. This means that at a later step $k > j$ in the derivation (after the

[10] Recall that the grounded extension is unique and, thus, credulous and skeptic inference over the grounded extension coincide (Dung, 1995).

revision process), the status of s changed to $s[\mathsf{e}]$. There are two options, either (i) at a later derivation step $k > j$ a (retrospective-)attack rule was applied to change the annotation of s or (ii) at some step $k > j$ the annotation of s was changed during another revision process.

Consider (i). In that case, a revision process is initiated at step k with $s \in \mathsf{RevSeq}$. Since the most recent annotation of r occurs at derivation step i, we know that r does not occur at any derivation step $k > i$. Consequently, $r \notin \mathsf{RevSeq}$. However, since s attacks r, $s \in \mathsf{RevSeq}$, and $r \notin \mathsf{RevSeq}$, the revision process triggers the application of **React** (on s and r) and so introduces the sequent $r[\mathsf{i}]$ at a derivation step $k > j \geq i$. This contradicts the assumption that the most recent annotation of $r[\mathsf{e}]$ occurs at i.

Consider (ii). Then, the status of s was changed during some revision process. In that case too, $s \in \mathsf{RevSeq}$, but $r \notin \mathsf{RevSeq}$ and so we reach a contradiction by reasoning similar to (i).

Retro This means that $s[\mathsf{e}]$ is a condition of **Retro** at derivation step i. By the side-condition on **Retro**, we know that $s[\mathsf{i}]$ is derived at derivation step $j > i$. Suppose that the most recent annotated version of s is not $[\mathsf{i}]$ or $[!]$. This means that $s[\mathsf{e}]$ was derived at some step $l > j$ after the revision process initiated at i (by the coherence of $\mathcal{D}$): again, either through a derivation step in which a (retrospective-)attack was applied or by another revision process. In both cases, we proceed as for the cases (i) and (ii) of the previous item and obtain a contradiction.

(2) For the last case, if the derivation step i was part of a revision process, this means that the annotation of s was revised during the process at some previous step $j < i$ and so $s \in \mathsf{RevSeq}$. Furthermore, at i also $r \in \mathsf{RevSeq}$ and so r and s are not revised further down the revision process. Then, to prove that the most recent annotation of s is still $[\mathsf{i}]$ or $[!]$ we suppose towards a contradiction that the status of s changed to $[\mathsf{e}]$. This must have occurred at a step $k > i$ after the revision process of which i was part. Then, proceed in accordance with the two cases of item (1).

Hence, $s \in \mathsf{Accept}(\mathcal{D})$. QED

Corollary 7.2 (Credulous inference and acceptability $[\mathsf{i}]$)**.** *Let $\mathcal{D}$ be a coherent $\mathcal{S}$-based* AC-*derivation and let $\Delta \Rightarrow \varphi \in \mathsf{Accept}(\mathcal{D})$ (i.e., $\mathcal{S} \mathrel{|\!\sim^{c}} \varphi$), then $\mathcal{AF}(\mathcal{D}) \mathrel{|\!\sim^{c}_{\mathtt{Stable}}} \varphi$.*

Remark 7.2. *The coherence requirement in Theorem 7.8 is necessary. To illustrate, consider the derivation $\mathcal{D}$ in Figure 7.3. Step 3 in Stage II is not coherent since the attacker s_3 in the derivation step initiating the revision process is eliminated at the end of the revision. In that case,* $\mathsf{Accept}(\mathcal{D}) = \{s_2\}$*, although $\mathcal{AF}(\mathcal{D})$ does not have a stable extension.*

Example 7.6. *In Example 7.5, if the last (retrospective-)attack of $p \Rightarrow p$ on $\neg p \Rightarrow \neg p$ is performed after the last (retrospective-)attack of $\neg p \Rightarrow \neg p$ on $p \Rightarrow p$, then for the* *corresponding derivation $\mathcal{D}$ we have $\mathsf{Accept}(\mathcal{D}) = \{\Rightarrow p \vee \neg p,\ q \Rightarrow q,\ p \Rightarrow p\}$. This is indeed a stable extension of $\mathcal{AF}(\mathcal{D})$. If the mutual attacks of $p \Rightarrow p$ and $\neg p \Rightarrow \neg p$ are performed in a reversed order, then $\mathsf{Accept}(\mathcal{D}) = \{\Rightarrow p \vee \neg p,\ q \Rightarrow q,\ \neg p \Rightarrow \neg p\}$, which again is a stable extension of $\mathcal{AF}(\mathcal{D})$.*

We now show a correspondence between the set $\mathsf{Final}(\mathcal{D})$ of finally derived sequents in $\mathcal{D}$ and the grounded extension of the argumentation framework induced by $\mathcal{D}$. We need the following definition.

Definition 7.12. *An $\mathcal{S}$-based derivation $\mathcal{D}$ is* saturated *if $\mathsf{Final}(\mathcal{D})$ is exhaustive in $\mathcal{D}$, i.e., the final acceptability rules are applied to every derived sequent in $\mathcal{D}$ to which it can be applied.*

We point out that saturation is a decidable property for a finite set $\mathcal{S} \subseteq \mathcal{L}$ of assertions. Namely, since an $\mathcal{S}$-based derivation is a finite sequence of tuples, it contains finitely many sequents. For each sequent, we can determine all its (finitely many) attackers based on $\mathcal{S}$. The following proposition shows the close relation between acceptable sequents in AC-derivations $\mathcal{D}$ and skeptic inference over stable semantics of argumentation frameworks induced by $\mathcal{D}$.

Proposition 7.9. *If an $\mathcal{S}$-based derivation $\mathcal{D}$ is saturated, then $\mathsf{Final}(\mathcal{D})$ is the (unique) grounded extension $\mathcal{E}$ of $\mathcal{AF}(\mathcal{D})$.*

Proof. We show that $\mathsf{Final}(\mathcal{D}) = \mathcal{E}$. Let $\langle s_1[!], \dots, s_n[!]\rangle$ be the ordered set of all $[!]$-annotated sequents derived in $\mathcal{D}$, in the order in which they occur in $\mathcal{D}$.

Left-to-Right. We prove by an induction on i that $s_i[!] \in \mathcal{E}$ for $i = 1, \dots, n$. *Base case.* Since $s_1[!]$ is derived by an application of a final applicability rule, and it is the first sequent in $\mathcal{D}$ with this property, it has no attackers. Since $\mathcal{E}$ is complete, $s_1 \in \mathcal{E}$. *Inductive step.* Suppose the sequent $s_{i+1}[!]$ was derived by a final acceptability rule calling upon the $[!]$-annotated sequents $s_{j_1}[!], \dots, s_{j_m}[!]$. Then, $j_1, \dots, j_m < i+1$ and by the inductive hypothesis, $s_{j_1}, \dots, s_{j_m} \in \mathcal{E}$. Also, the sequents $s_{j_1}, \dots, s_{j_m}$ attack all attackers of s_{i+1}, so s_{i+1} is defended by $\mathcal{E}$ and by the completeness of $\mathcal{E}$, $s_{i+1} \in \mathcal{E}$.

Right-to-Left. We show that $\mathsf{Final}(\mathcal{D})$ is complete in $\mathcal{AF}(\mathcal{D})$. Since $\mathcal{E}$ is $\subseteq$-minimal complete, then $\mathcal{E} \subseteq \mathsf{Final}(\mathcal{D})$. We show conflict-freeness inductively by showing that for each $i = 1, \dots, n$ there is no $k \leq i$ such that $s_k \in \mathsf{Final}(\mathcal{D})$ attacks $s_i \in \mathsf{Final}(\mathcal{D})$. *Base case.* Trivial, since s_1 does not have any attackers. *Inductive step.* Suppose there is a $k \leq i$ s.t. s_k attacks s_{i+1}. By the final acceptability rule, there is a $j \leq i$ s.t. s_j attacks s_k, which contradicts the inductive hypothesis. Suppose now that s_{i+1} attacks itself. By the final acceptability rule, there is a $k \leq i$ such that s_k attacks s_{i+1}, which we have already excluded.

For admissibility, suppose that $s \in \mathsf{Derived}(\mathcal{D}) \setminus \mathsf{Final}(\mathcal{D})$ attacks some $s_i \in \mathsf{Final}(\mathcal{D})$. Then by the application of the final acceptability rule that produces $s_i[!] \in \mathcal{D}$, there is a

$k < i$ such that s_k attacks s with $s_k \in \mathsf{Final}(\mathcal{D})$. For completeness, let $s \in \mathsf{Derived}(\mathcal{D})$ be defended by $\mathsf{Final}(\mathcal{D})$. Then we can apply the final applicability rule to derive $s[!]$. Since $\mathcal{D}$ is saturated, $s[!] \in \mathsf{Final}(\mathcal{D})$. QED

Corollary 7.3 (Skeptic inference and final acceptability [!])**.** *Let $\mathcal{D}$ be a saturated $\mathcal{S}$-based derivation, let $\Delta \Rightarrow \varphi \in \mathsf{Final}(\mathcal{D})$ (i.e., $\mathcal{S} \mathrel{|\!\sim^{s}} \varphi$), then $\mathcal{AF}(\mathcal{D}) \mathrel{|\!\sim^{s}_{\texttt{Grounded}}} \varphi$.*

The final proposition expresses that for each AC-sequent in the grounded extension of an argumentation framework, we can construct a AC-derivation in which the sequent is derived as finally acceptable.

Proposition 7.10. *For a derivation $\mathcal{D}$, let $\mathcal{S} = \bigcup\{\Delta \mid \Delta \Rightarrow \varphi \in \mathcal{D}\}$. Let $\mathcal{AF}(\mathcal{S}) = \langle \mathsf{Arg}(\mathcal{S}), \mathsf{Att}\rangle$ where $\mathsf{Arg}(\mathcal{S}) = \{\Gamma \Rightarrow \psi \mid \Gamma \subseteq \mathcal{S}$ and $\Gamma \vdash \psi\}$ and $\mathsf{Att} = \{(s,t) \mid s = \Delta \Rightarrow \varphi, t = \Gamma \Rightarrow \psi \in \mathsf{Arg}$ and $\varphi \vdash \neg\bigwedge\Gamma\}$. Let $\mathcal{E}$ be the grounded extension of $\mathcal{AF}(\mathcal{S})$. For every $s \in \mathcal{E}$, there is an $\mathcal{S}$-based derivation $\mathcal{D}'$ of $s[!]$ that does not contain applications of the rules **Def**, **React**, and **Retro**.*

Proof. We note that for each $r \in \mathsf{Arg}(\mathcal{S})$ there is an $\mathcal{S}$-based derivation $\mathcal{D}_r$ concluding $r[i]$. Since $\mathcal{D}$ is finite, we know the grounded extension can be characterized as follows: $\mathcal{E} = \bigcup_{i\geq 1} \mathcal{E}_i$ where $\mathcal{E}_0 = \emptyset$ and $\mathcal{E}_{i+1} = \{s \in \mathsf{Arg} \mid \mathcal{E}_i \text{ defends } s\}$ (Dung, 1995). Let $s \in \mathcal{E}$. We prove by induction on i, that for each $s \in \mathcal{E}_i$ $(i = 1, \ldots, n)$ there is an $\mathcal{S}$-based derivation $\mathcal{D}'_s$ of $s[!]$ without attacks. *Base case.* Since $s \in \mathcal{E}_1$, s has no attackers. Let $\mathcal{D}'_s$ be the extension of $\mathcal{D}_s$ by an application of the final acceptability rule that concludes $s[!]$. *Inductive step.* Let $s \in \mathcal{E}_{i+1}$. For each attacker r of s there is an $s' \in \bigcup_{j=1}^{i} \mathcal{E}_j$ that attacks r. By the IH there are derivations $\mathcal{D}'_{s'}$ for each such defending s'. We obtain $\mathcal{D}'_s$ by concatenating the proofs $\mathcal{D}_r$ of each attacker, the proofs $\mathcal{D}'_{s'}$ for each defender, and applying the final acceptability rule to conclude $s[!]$. QED

In fact, a stronger result than Proposition 7.10 is possible, where a *coherent* derivation is constructed (without using the rules **React** and **Retro**). This derivation can be straightforwardly obtained since each attacker of the grounded extension $\mathcal{E}$ is counterattacked, and thus eliminable, by a finally acceptable sequent from $\mathcal{E}$.

7.5 Defeasible Normative Reasoning and Annotated Calculi

This section addresses this chapter's third and last goal (Objective 3). Namely, we extend the monotonic Deontic Argumentation Calculi DAC from Chapter 6 with annotations and annotation revision rules.[11] We call the resulting proof systems Annotated Deontic Argumentation Calculi (for short, ADAC). We prove that the consequence relation of ADAC is nonmonotonic. We provide some first results by proving a correspondence

[11] Applications of annotated calculi to other types of base logics are likewise possible, e.g., think of applications to modal logics such as epistemic logics. Such applications must be left to future work.

$$\frac{\Gamma_1, (\varphi, \psi) \Rightarrow^{[\mathsf{i}]} \Delta \qquad \Gamma_2 \Rightarrow^{[\mathsf{i}]} \neg(\varphi, \psi)}{\Gamma_1, (\varphi, \psi) \Rightarrow^{[\mathsf{e}]} \Delta}\ \mathbf{Def}_x$$

$$\frac{\Gamma_1, (\varphi, \psi) \Rightarrow^{[\mathsf{e}]} \Delta \qquad \Gamma_2 \Rightarrow^{[\mathsf{e}]} \neg(\varphi, \psi)}{\Gamma_1, (\varphi, \psi) \Rightarrow^{[\mathsf{i}]} \Delta}\ \mathbf{React}_x$$

$$\begin{array}{c}\dfrac{\Gamma_1, (\varphi, \psi) \Rightarrow^{[\mathsf{i}]} \Delta \qquad \Gamma_2 \Rightarrow^{[\mathsf{e}]} \neg(\varphi, \psi)}{\Gamma_1, (\varphi, \psi) \Rightarrow^{[\mathsf{e}]} \Delta}\ \mathbf{Retro}_x \\ \vdots \\ \dfrac{\Gamma_2 \Rightarrow^{[\mathsf{e}]} \neg(\varphi, \psi) \qquad\qquad \Gamma_3 \Rightarrow^{[\mathsf{e}]} \neg(\theta, \chi)}{\Gamma_2 \Rightarrow^{[\mathsf{i}]} \neg(\varphi, \psi)}\ \mathbf{React}_x^a\end{array}$$

$$\frac{\begin{array}{ll}\text{(i)} & \Gamma_1 \Rightarrow^{[\mathsf{i}]} \Delta_1 \\ \text{(ii)} & (\forall \Gamma_2 \Rightarrow \Delta_2 \in \mathsf{Def}(\Gamma_1))\ \Gamma_2 \Rightarrow^{[*]} \Delta_2 \\ \text{(iii)} & (\forall \Gamma_2 \Rightarrow \Delta_2 \in \mathsf{Def}(\Gamma_1),\ \exists \Gamma_3 \Rightarrow \Delta_3 \in \mathsf{Def}(\Gamma_2))\ \Gamma_3 \Rightarrow^{[!]} \Delta_3\end{array}}{\Gamma_1 \Rightarrow^{[!]} \Delta_1}\ \mathbf{Final}_x$$

Figure 7.5: The annotation revision rules $\mathbf{Def}_x$, $\mathbf{React}_x$, $\mathbf{Retro}_x$, and $\mathbf{Final}_x$ of ADAC with the the underlying base calculus $\mathsf{DAC}_\mathcal{S}$. In $\mathbf{Final}_x$, $\mathsf{Def}(\Gamma) = \{\Delta \Rightarrow \neg(\varphi, \psi)$ is $\mathsf{DAC}_\mathcal{S}$-derivable $\mid \Delta \subseteq \Sigma$ and $(\varphi, \psi) \in \Gamma\}$ for a finite set $\Sigma \subseteq \mathcal{L}^{io}$. The side-condition (a) stipulates that $(\theta, \chi) \in \Gamma_2$. The rule $\mathbf{Retro}_x$ is a *system of rules*, which stipulates that each application of the topmost rule must be followed by an application of a corresponding $\mathbf{React}_x$ rule during the $\mathbf{Retro}_x$ initiated revision process. For representational reasons, the conditions (i)-(iii) of $\mathbf{Final}_x$ are given vertically.

between the inference relation of ADAC and specific semantic extensions of argumentation frameworks. Last, we extend ADAC with the annotation $[\perp]$ and a corresponding rule capturing the notion of finally eliminated sequents.

First, we point out that the language of DAC differs from the ones employed for AC. It allows us to simplify some of the annotation revision rules (Definition 7.5). Second, DAC also assume underlying base logics and corresponding sound and complete calculi. In fact, the requirements imposed on the base logics for DAC satisfy all requirements for that of AC (see Remark 7.3 below). Last, throughout this section, we highlight various future work directions. In what follows, we will not recall the definitions of DAC but refer where necessary to Chapter 6 for details.

Remark 7.3. *Comparing Definition 6.2 (Chapter 6, page 220) and Definition 7.2 it can be straightforwardly checked that the base logics underlying* DAC *belong to the class of base logics defined for* AC.

Definition 7.13 (The Labelled Input/Output Language Definition 6.6 (Chapter 6))**.** *Let $\mathcal{L}^i$ with $i \in \{f, o, c\}$ be defined through the following BNF grammar:*

$$\varphi^i ::= p^i \mid \top^i \mid \bot^i \mid (\neg\varphi)^i \mid (\varphi \wedge \varphi)^i \mid (\varphi \vee \varphi)^i \mid (\varphi \rightarrow \varphi)^i$$

with $p \in \mathsf{Atoms}$. Let $\mathcal{L}^{\downarrow}$ be the language $\mathcal{L}^i$ stripped from its labels. Let $\mathcal{L}^n = \{(\varphi, \psi) \mid \varphi, \psi \in \mathcal{L}^{\downarrow}\}$. The language of norms *is defined as $\mathcal{L}^n \cup \overline{\mathcal{L}^n}$, where $\overline{\mathcal{L}^n} = \{\neg(\varphi, \psi) \mid (\varphi, \psi) \in \mathcal{L}^n\}$ is the language expressing the inapplicability of norms. Let $\mathcal{L}^{io} = \mathcal{L}^f \cup \mathcal{L}^o \cup \mathcal{L}^c \cup \mathcal{L}^n \cup \overline{\mathcal{L}^n}$ be the* labelled I/O language.

Recall that the labels f, o, and c, express *facts, obligations,* and *constraints*, respectively. In what follows, we directly restrict sequents to a given knowledge base $\mathcal{K} = \langle \mathcal{F}, \mathcal{N}, \mathcal{C} \rangle$ where $\mathcal{F} \subseteq \mathcal{L}^f$ represent the factual context, $\mathcal{N} \subseteq \mathcal{L}^n$ a normative code, and $\mathcal{C} \subseteq \mathcal{L}^c$ a set of constraints with which the generated output has to be consistent. Furthermore, it suffices for our present aims to take any derivable DAC sequents as initial sequents of ADAC. This means that the rules of DAC are not directly incorporated in ADAC.

To facilitate readability, we recall the two types of *dac* arguments. First, we have arguments that conclude obligations on the basis of facts and norms from $\mathcal{K}$. For instance,

$$\varphi^f, (\varphi, \psi), (\psi, \theta) \Rightarrow \theta^o$$

where $\varphi^f \in \mathcal{F}$, $(\varphi, \psi), (\psi, \theta) \in \mathcal{N}$ (and in this case, the calculus in question contains the rule **L-CT**). The left-hand side of the sequent constitutes the reasons for the obligation θ. Second, we have arguments that conclude the *inapplicability of norms* in a given context. For instance,

$$\varphi^f, (\varphi, \psi), \neg\theta^c \Rightarrow \neg(\psi, \theta)$$

where $\varphi^f \in \mathcal{F}$, $(\varphi, \psi) \in \mathcal{N}$, and $\neg\theta^c \in \mathcal{C}$. The latter type of arguments provide the attacks in ADAC. Recall that this type of attack expresses an *undercut*, namely, an attack on a reason (ψ, θ) provided in a sequent's premises. Since such attacking arguments directly provide their own attacking conditions, we can simplify the annotation revision rules for ADAC (see Figure 7.5). In Definition 7.14, we define the calculi. We briefly discuss the rules.

The rule $\mathbf{Def}_x$ represents a straightforward adaptation of the notion of attack between DAC arguments, namely, a sequent s_1 attacks some other sequent s_2 whenever the former concludes the inapplicability of a norm used as a premise in the latter. The rule $\mathbf{React}_x$ reinstates sequents whose attacker has been eliminated. The rule $\mathbf{Retro}_x$ allows for eliminated attackers provided that the attacker is reinstated further down in the derivation. The final acceptability rule $\mathbf{Final}_x$ deserves some more discussion. The set $\mathsf{Def}(\Gamma)$ is the collection of all DAC-derivable arguments concluding the inapplicability of some norm (φ, ψ) occurring in Γ. In the case that $\Gamma \cap \mathcal{L}^n = \emptyset$ the set $\mathsf{Def}(\Gamma)$ is trivially empty. Consequently, the second condition of $\mathbf{Final}_x$ collects all attackers of the sequent whose final acceptability is to be determined. The third condition specifies for each of these attackers a finally accepted attacker counter-attacking it. For this rule to be well-defined, we restrict Def to a finite set $\Sigma \subseteq \mathcal{L}^{io}$ (e.g., a finite knowledge base $\mathcal{K}$, see the footnote on page 262). In order to obtain more concise derivations, we require that the employed DAC arguments satisfy the relevance condition defined in Section 6.8. In the sequel, we leave this assumption implicit.

Definition 7.14. *Let $\mathsf{DAC}^{\mathsf{r}}_{\mathcal{S}}$ be a relevance-aware Deontic Argumentation Calculus from Definition 6.13 (in what follows, we omit the index r). Let $\mathcal{K} = \langle \mathcal{F}, \mathcal{N}, \mathcal{C} \rangle$ be a finite knowledge base. An* Annotated Deontic Argumentation Calculus *(for short,* $\mathsf{ADAC}_{\mathcal{S}}$*) consists of the following components.*

- ***Annotated initial rules*** *for each $\mathsf{DAC}_{\mathcal{S}}$-derivable sequents $\Gamma \Rightarrow \Delta$ where each initial sequent is labelled* $[\mathsf{i}]$*;*
- ***The annotation revision rules Def$_x$, React$_x$, Retro$_x$,*** *and **Final$_x$** from Figure 7.5.*

A $\mathcal{K}$-based derivation *$\mathcal{D}$ of $\mathsf{ADAC}_{\mathcal{S}}$ is a finite sequence of tuples $\mathcal{D} = \langle T_1, \dots, T_n \rangle$ where the index $1 \leq i \leq n$ determines the tuple's order in the derivation $\mathcal{D}$. Each T_i contains the derived annotated sequent (the tuple's sequent), the derivation rule that is applied (the tuple's rule), and the indexes of the tuples whose sequents serve as the conditions of the tuple's rule. Each sequent $\Gamma \Rightarrow^{[\mathsf{a}]} \Delta$ (with $a \in \{i, e, !\}$) is $\mathcal{K}$-*dependent*, that is, $\Gamma \subseteq \mathcal{F} \cup \mathcal{N} \cup \mathcal{C}$.*[12]

We impose the following restrictions on $\mathcal{K}$-based derivations:

- *During the construction of a derivation $\mathcal{D}$, after each extension with a tuple T_i containing an attack rule (**Def$_x$**) or a retrospective attack rule (**Retro$_x$**), an* annotation revision process *is initiated, and the derivation is extended with new attack or reactivation rules for updating the sequent annotations when necessary.*
- *Reactivation rules are applied only during a revision process.*

The revision process *is the one described in Definition 7.7, the notion of* most recent status *is that of Definition 7.6, and the notion of* coherence *is that of Definition 7.8.*

In what follows, we focus on *coherent* ADAC-derivations only (Definition 7.14). Furthermore, since ADAC does not contain DAC rules—taking DAC sequents as initial sequents—one may think of ADAC as a *postliminary* proof-theoretic approach for resolving normative conflicts in DAC; cf. (Bonatti and Olivetti, 2002).

Definition 7.15 (Skeptic and credulous inference ADAC)**.** *For an annotated deontic argumentation calculus $\mathsf{ADAC}_{\mathcal{S}}$, we define skeptical (s) and credulous (c) inference as follows:*

[12] In the case where a derivation $\mathcal{D}$ involves applications of **Final$_x$** we require $\mathcal{K}$ to be finite.

- *$\mathcal{K} \mathrel{|\!\!\sim}^{s}_{\mathsf{ADAC}_{\mathcal{S}}} \psi$ if there is a $\mathcal{K}$-based $\mathsf{ADAC}_{\mathcal{S}}$-derivation $\mathcal{D}$ such that $\Gamma \Rightarrow^{[\mathsf{!}]} \psi$ is derived in $\mathcal{D}$.*
- *$\mathcal{K} \mathrel{|\!\!\sim}^{c}_{\mathsf{ADAC}_{\mathcal{S}}} \psi$ if there is a $\mathcal{K}$-based $\mathsf{ADAC}_{\mathcal{S}}$-derivation $\mathcal{D}$ such that $\Gamma \Rightarrow^{[i]} \psi$ is derived in $\mathcal{D}$ and $[i]$ is the sequent's most recent status.*

Henceforth, reference to the subscript $\mathsf{ADAC}_{\mathcal{S}}$ is omitted.

First, we observe that the following propositions that hold for AC are preserved in the context of ADAC.

Proposition 7.11. *Let $\mathcal{D}$ be a coherent $\mathcal{K}$-based $\mathsf{ADAC}_{\mathcal{S}}$-derivation with a finite $\mathcal{K}$ and let $s[\mathsf{!}] \in \mathcal{D}$. The most recent status of s is either $[\mathsf{!}]$ or $[\mathsf{i}]$.*

Proof. It can be straightforwardly observed that the proof is identical to that of Proposition 7.3. QED

Corollary 7.4. *Let $\mathcal{D}$ be a coherent $\mathcal{K}$-based $\mathsf{ADAC}_{\mathcal{S}}$-derivation with a finite $\mathcal{K}$ and let s attack r somewhere in $\mathcal{D}$ such that $s[\mathsf{!}] \in \mathcal{D}$. The most recent status of r is $[\mathsf{e}]$.*

Next, we prove that the consequence relation $\mathrel{|\!\!\sim}^{s}_{\mathsf{ADAC}_{\mathcal{S}}}$ satisfies certain desirable properties (cf. Section 7.3). Proposition 7.12 expresses that $\mathsf{ADAC}_{\mathcal{S}}$ preserves derivability in $\mathsf{DAC}_{\mathcal{S}}$ whenever a knowledge base is consistent. Proposition 7.13 expresses that the class of ADAC is paraconsistent.

Proposition 7.12. *Let $\mathsf{ADAC}_{\mathcal{S}}$ be an annotated deontic argumentation calculus based on $\mathsf{DAC}_{\mathcal{S}}$ of Definition 6.8 (Section 6.1). Let $\mathcal{K} = \langle \mathcal{F}, \mathcal{N}, \mathcal{C} \rangle$ be a knowledge base. We say $\mathcal{K} \vdash_{\mathsf{DAC}_{\mathcal{S}}} \varphi^{o}$ if there exists an $\mathsf{DAC}_{\mathcal{S}}$-derivable sequent $\Gamma \Rightarrow \varphi^{o}$ with $\Gamma \subseteq \mathcal{K}$. If $\mathcal{K}$ is consistent, that is, $\mathcal{N}$ is the only maximal-consistent set of norms according to Definition 6.5 (Section 6.1), then $\mathcal{K} \vdash_{\mathsf{DAC}_{\mathcal{S}}} \varphi^{o}$ iff $\mathcal{K} \mathrel{|\!\!\sim}^{s}_{\mathsf{ADAC}_{\mathcal{S}}} \varphi^{o}$.*

Proof. **Left-to-Right.** Since $\mathcal{K}$ is consistent, this means that no argument $\Gamma \Rightarrow \neg(\varphi, \psi)$ is derivable such that $\Gamma \subseteq \mathcal{K}$ and $(\varphi, \psi) \in \mathcal{N}$. This means there are no $\mathcal{K}$-based attacking sequents $\mathsf{DAC}_{\mathcal{S}}$-derivable. Hence, for each $\mathsf{DAC}_{\mathcal{S}}$-derivable $\Gamma \Rightarrow \varphi^{o}$, we can construct a $\mathcal{K}$-based $\mathsf{ADAC}_{\mathcal{S}}$-derivation that contains an application of $\mathbf{Final}_{x}$ to $\Gamma \Rightarrow \varphi^{o}$ and so $\Gamma \Rightarrow^{[\mathsf{!}]} \varphi^{o}$ is derivable. Hence, $\mathcal{K} \mathrel{|\!\!\sim} \varphi^{o}$. **Right-to-Left.** Trivial. QED

Proposition 7.13. *Let $\mathsf{ADAC}_{\mathcal{S}}$ be an annotated deontic argumentation calculus based on $\mathsf{DAC}_{\mathcal{S}}$. Let the knowledge base $\mathcal{K}$ be $\mathcal{F} = \emptyset$, $\mathcal{N} = \{(\top, p), (\top, \neg p)\}$ and $\mathcal{C} = \emptyset$. Then, ADAC is paraconsistent, i.e., $\mathcal{K} \mathrel{|\!\!\not\sim} q^{o}$.*

Proof. Among others, we have the following relevant $\mathsf{DAC}_{\mathcal{S}}$-derivable sequents as initial $\mathsf{ADAC}_{\mathcal{S}}$ sequents: $a = \top, (\top, p) \Rightarrow^{[\mathsf{i}]} p^{o}$, $b = \top, (\top, \neg p) \Rightarrow^{[\mathsf{i}]} \neg p^{o}$, $c = \top, (\top, p), (\top, \neg p) \Rightarrow^{[\mathsf{i}]} q^{o}$, $d = \top, (\top, p) \Rightarrow^{[\mathsf{i}]} \neg(\top, \neg p)$, and $e = \top, (\top, \neg p) \Rightarrow^{[\mathsf{i}]} \neg(\top, p)$. Even though argument

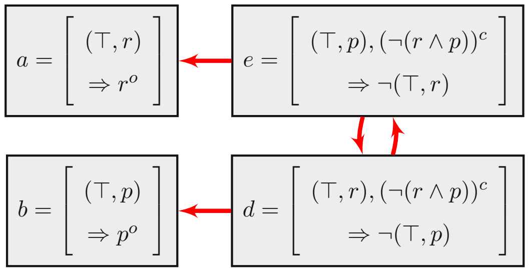

$a = \left[\begin{array}{c}(\top, r)\\ \Rightarrow r^o\end{array}\right]$
$e = \left[\begin{array}{c}(\top, p), (\neg(r \wedge p))^c\\ \Rightarrow \neg(\top, r)\end{array}\right]$
$b = \left[\begin{array}{c}(\top, p)\\ \Rightarrow p^o\end{array}\right]$
$d = \left[\begin{array}{c}(\top, r), (\neg(r \wedge p))^c\\ \Rightarrow \neg(\top, p)\end{array}\right]$

Figure 7.6: Defeasible normative reasoning examples: A deontic conflict (Example 7.7). Argument e defends $\{b, e\}$, whereas argument d defends $\{a, d\}$.

c concludes q^o from the inconsistent $\mathcal{K}$, the reasons used to conclude q^o are both attacked by arguments d and e. Since d and e reciprocally attack each other and are the only attacking arguments, we have that neither of these attackers can be finally derived. Therefore c cannot be finally derived either. QED

Proposition 7.14. *The consequence relation $\mid\!\sim$ of* ADAC *is nonmonotonic.*

Proof. It suffices to provide an example. Consider the knowledge base $\mathcal{K}$ consisting of $\mathcal{F} = \emptyset$, $\mathcal{N} = \{(\top, p), (q, \neg p)\}$, and $\mathcal{C} = \emptyset$. Clearly, $\mathcal{K} \mid\!\sim p^o$ since the argument $\top^f, (\top, p) \Rightarrow p^o$ cannot be attacked relative to $\mathcal{K}$. A proper extension $\mathcal{K}'$ of $\mathcal{K}$ with either $\mathcal{F} = \{q^f\}$, $\mathcal{N} = \{(\top, p), (q, \neg p), (\top, \neg p)\}$, or $\mathcal{C} = \{\neg p^c\}$ would result in $\mathcal{K}' \not\mid\!\sim p^o$. QED

To illustrate the use of ADAC, we reconsider the deontic dilemma from Chapter 6, page 218.

Example 7.7 (Deontic Dilemma, Figure 7.6)**.** *Joan has an obligation to* return *a borrowed hammer to her friend Maxwell* $(\top, r)$*. Furthermore, Joan knows Max is planning to commit a crime with the hammer, and she is under the obligation to* prevent *harm from being done* $(\top, p)$*. Furthermore, the constraint is that Joan cannot secure both r and p, i.e., $\neg(r \wedge p)^c$. Joan is in a deontic dilemma. The knowledge base $\mathcal{K}$ is $\mathcal{F} = \emptyset$, $\mathcal{N} = \{(\top, r), (\top, p)\}$, and $\mathcal{C} = \{\neg(r \wedge p)^c\}$. We assume the underlying base logic to be classical. The* DAC *arguments that can be constructed are presented in Figure 6.2. These serve as initial sequents in* ADAC*. The two defeating arguments, d and e, express that given the constraints, one of either two norms cannot be asserted.*

The following coherent ADAC$_{\mathcal{S}}$*-derivation (with* $\boldsymbol{C\text{-}T} \in \mathcal{S}$*) expresses the conflict, constituting a stable set $\{e\}$ for the argument e justifying that Joan is not under obligation to return the hammer.*

$$\dfrac{a[\mathsf{i}] \qquad \dfrac{\dfrac{e[\mathsf{i}] \quad d[\mathsf{i}]}{e[\mathsf{e}]}\,\boldsymbol{Def}_x \qquad \dfrac{d[\mathsf{i}] \quad e[\mathsf{i}]}{d[\mathsf{e}]}\,\boldsymbol{Def}_x}{e[\mathsf{i}]}\,\boldsymbol{React}_x}{a[\mathsf{e}]}$$

The derivation can be coherently extended with the following rule applications:

$$\dfrac{\begin{array}{c}\vdots\\ a[\mathsf{e}]\end{array} \qquad \dfrac{e[\mathsf{i}] \qquad d[\mathsf{i}]}{e[\mathsf{e}]}\,\boldsymbol{Def}_x}{a[\mathsf{i}]}\,\boldsymbol{React}_x$$

The resulting extension constitutes the stable set of arguments $\{d, a\}$ justifying Joan's returning of the hammer. Observe that similar derivations can be obtained for the argument b. However, the derivation cannot be extended in a tree-like manner in order to include argument b. We come back to this below in Example 7.8. In both cases, we have neither $\mathcal{K} \mid\!\sim^s r^o$ nor $\mathcal{K} \mid\!\sim^s p^o$, as desired.

7.5.1 Relations to Logical Argumentation

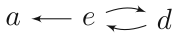
We show that the correspondence between AC and argumentation frameworks is preserved in the case of ADAC.

Definition 7.16 (Argumentation Frameworks induced by ADAC)**.** *Let $\mathsf{ADAC}_{\mathcal{S}}$ be an annotated deontic argumentation calculus, let $\mathcal{K}$ be a knowledge base, and let $\mathcal{D}$ be a $\mathcal{K}$-based $\mathsf{ADAC}_{\mathcal{S}}$-derivation. Then:*

- $\mathsf{Derived}(\mathcal{D})$ *is the set of sequents* s *s.t.* $s[\mathsf{i}] \in \mathcal{D}$*;*
- $\mathsf{Accept}(\mathcal{D})$ *is the set of sequents* s *in* $\mathsf{Derived}(\mathcal{D})$ *such that their most updated status is* $[\mathsf{i}]$ *or* $[!]$*;*
- $\mathsf{Final}(\mathcal{D})$ *is the set of sequents s in* $\mathsf{Derived}(\mathcal{D})$ *such that* $s[!] \in \mathcal{D}$*;*
- $\mathsf{Att}(\mathcal{D})$ *is the set of pairs (s_1, s_2) such that s_1 attacks s_2 by an application of either* $\boldsymbol{Def}_x$*,* $\boldsymbol{React}_x$*,* $\boldsymbol{Retro}_x$*, or* $\boldsymbol{Final}_x$ *in $\mathcal{D}$, with $s_1, s_2 \in$* $\mathsf{Derived}(\mathcal{D})$*.*

$\mathcal{AF}(\mathcal{D}) = \langle \mathsf{Derived}(\mathcal{D}), \mathsf{Att}(\mathcal{D}) \rangle$ *is called the (sequent-based) argumentation framework induced by $\mathcal{D}$. The definitions of* semantic *extensions of jointly acceptable arguments are as in Definition 7.11.*

Example 7.8 (Deontic Dilemma, Example 7.7 cont.)**.** *Reconsider the* $\mathsf{ADAC}_{\mathcal{S}}$*-derivation in Example 7.7. There we saw that the derivation could not be extended with argument b while simultaneously preserving the tree-like structure of the derivation. The resulting argumentation framework from that derivation is as follows:*



In fact, the stable extensions $\{a, d\}$ *and* $\{e\}$ *correspond to the related admissible sets* $\mathcal{E}_a$ *and* $\mathcal{E}_e$*, respectively. Recall that a related admissible set* $\mathcal{E}_a$ *identifies the relevant arguments that justify the acceptability of* a*. There is an interesting correspondence between tree-like* ADAC*-derviations and related admissibility. The tree-like structure enforces that each argument occurring in the tree-like derivation* $\mathcal{D}$ *is* indirectly related *to all other arguments in* $\mathcal{D}$ *by means of attack relations. We leave a formal investigation for future work.*

It can be observed that the correspondence results for AC are preserved in the context of ADAC. The reason is that the proofs of Section 7.4 depend on the definition of most recent status (Definition 7.6) and the definition of the revision process (Definition 7.7), which in both cases are the same as for AC. The only difference is that the annotation revision rules of ADAC do not contain the (third) attacking condition. However, it can also be observed that the proofs do not depend on the rule's third condition.

Proposition 7.15. *Let* $\mathcal{D}$ *be coherent* $\mathcal{K}$*-based* $\mathsf{ADAC}_{\mathcal{S}}$*-derivation:* $\mathsf{Accept}(\mathcal{D})$ *is a stable extension of* $\mathcal{AF}(\mathcal{D})$*.*

Proof. See proof of Proposition 7.8. QED

Corollary 7.5 (Credulous inference and acceptability [i])**.** *Let* $\mathcal{D}$ *be a coherent* $\mathcal{K}$*-based* $\mathsf{ADAC}_{\mathcal{S}}$*-derivation, let* $\Delta \Rightarrow \varphi \in \mathsf{Accept}(\mathcal{D})$ *(i.e.,* $\mathcal{K} \mathrel{|\!\sim^{c}} \varphi$*), then* $\mathcal{AF}(\mathcal{D}) \mathrel{|\!\sim^{c}_{\texttt{Stable}}} \varphi$*.*

Proposition 7.16. *If an* $\mathcal{K}$*-based* $\mathsf{ADAC}_{\mathcal{S}}$*-derivation* $\mathcal{D}$ *is saturated, then* $\mathsf{Final}(\mathcal{D})$ *is the (unique) grounded extension* $\mathcal{E}$ *of* $\mathcal{AF}(\mathcal{D})$*.*

Proof. See proof of Proposition 7.9. QED

Corollary 7.6 (Skeptic inference and final acceptability [!])**.** *Let* $\mathcal{D}$ *be a saturated* $\mathcal{K}$*-based* $\mathsf{ADAC}_{\mathcal{S}}$*-derivation, let* $\Delta \Rightarrow \varphi \in \mathsf{Final}(\mathcal{D})$ *(i.e.,* $\mathcal{K} \mathrel{|\!\sim^{s}} \varphi$*), then* $\mathcal{AF}(\mathcal{D}) \mathrel{|\!\sim^{s}_{\texttt{Grounded}}} \varphi$*.*

Proposition 7.17. *Let* $\mathsf{ADAC}_{\mathcal{S}}$ *be a calculus based on* $\mathsf{DAC}_{\mathcal{S}}$*. For a derivation* $\mathcal{D}$*, let* $\mathcal{K} = \bigcup\{\Delta \mid \Delta \Rightarrow \varphi^{o} \in \mathcal{D}\}$*. Let* $\mathcal{AF}(\mathcal{K}) = \langle \mathsf{Arg}(\mathcal{K}), \mathsf{Att} \rangle$ *where* $\mathsf{Arg}(\mathcal{K}) = \{\Gamma \Rightarrow \psi^{o} \mid \mathcal{K} \vdash_{\mathsf{DAC}_{\mathcal{S}}} \Gamma \Rightarrow \psi^{o}\}$ *and* $\mathsf{Att} = \{(s, t) \mid s = \Delta \Rightarrow \neg(\varphi, \psi), t = \Gamma, (\varphi, \psi) \Rightarrow \Sigma \in \mathsf{Arg}\}$*. Let* $\mathcal{E}$ *be the grounded extension of* $\mathcal{AF}(\mathcal{K})$*. For every* $s \in \mathcal{E}$*, there is an* $\mathcal{K}$*-based derivation* $\mathcal{D}'$ *of* $s[!]$ *without applications of the rules* $\boldsymbol{Def}_x$*,* $\boldsymbol{React}_x$*, and* $\boldsymbol{Retro}_x$*.*

Proof. See the proof of Proposition 7.10. QED

It remains an open question to determine the relation between maximally consistent sets generated by the class of constraint Input/Output logics in Chapter 6, the argumentation frameworks induced by DAC, and the annotated deontic argumentation calculi ADAC presented in this section. We come back to this in Section 7.6.

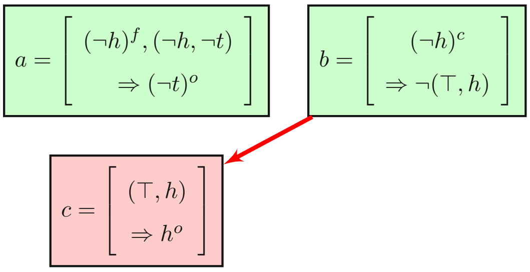

$a = \left[\begin{array}{c} (\neg h)^f, (\neg h, \neg t) \\ \Rightarrow (\neg t)^o \end{array} \right]$
$b = \left[\begin{array}{c} (\neg h)^c \\ \Rightarrow \neg(\top, h) \end{array} \right]$
$c = \left[\begin{array}{c} (\top, h) \\ \Rightarrow h^o \end{array} \right]$

Figure 7.7: The contrary-to-duty scenario of Example 7.9.

7.5.2 Finally Eliminable Arguments

In this last part, we consider some extensions of ADAC with additional rules that serve the explanatory aims underlying DAC (Chapter 6, page 212). First, we extend the set of labels with a label $[\bot]$, which expresses that a sequent is *finally eliminated*. As will be shown, such sequents can never occur in a stable extension of its corresponding argumentation framework. The identification of such arguments is relevant for explanations since they express DAC-derivable arguments containing combinations of norms that can never be jointly applicable. The rule $\mathbf{FinalElim}_x$ is defined as follows:

$$\frac{\Gamma_1, (\varphi, \psi) \Rightarrow^{[*]} \Delta \qquad \Gamma_2 \Rightarrow^{[!]} \neg(\varphi, \psi)}{\Gamma_1, (\varphi, \psi) \Rightarrow^{[\bot]} \Delta}\ \mathbf{FinalElim}_x$$

The rule functions like a regular attack rule $\mathbf{Def}_x$ but results in a more informative annotation. Intuitively, any argument attacked by a finally accepted argument is permanently eliminated. The revision process of ADAC must also be adjusted to incorporate $\mathbf{FinalElim}_x$. This is straightforward: it behaves like the clause for $\mathbf{Def}_x$. Furthermore, since $s[\bot]$ sequents are permanently eliminated, the rules $\mathbf{React}_x$ and $\mathbf{Retro}_x$ can be refined to allow only eliminated sequents $s[\mathsf{e}]$ in the first condition, respectively, the second condition of the rule such that there is no $s[\bot]$ occurring in the derivation prior to that rule's application.[13] Let us reconsider the contrary-to-duty scenario from Chapter 6, page 217.

Example 7.9 (Contrary-to-Duty Reasoning, Figure 7.7)**.** *Reconsider (a simplification of) the* Contrary-to-Duty *scenario from Example 6.1 of Section 6.1. Billy is obliged to go and help her neighbors* $(\top, h)$. *If Billy does not go, she ought not to tell them she goes* $(\neg h, \neg t)$. *Furthermore, Billy does not go to help her neighbors* $\neg h^f$. *Billy needs to know her obligations consistent with the fact that she does not go to help, i.e., given that she has violated the norm* $(\top, h)$. *Let the knowledge base* $\mathcal{K}$ *consist of* $\mathcal{F} = \{\neg h^f\}$ *and* $\mathcal{N} = \{(\top, h), (\neg h, \neg t)\}$, *and let the constraints be* $\mathcal{C} = \{\neg h^c\}$. *Figure 7.7 presents the* $\mathcal{K}$*-based* DAC*-derivable arguments a, b, and c. Given the CTD situation in which Billy resides, she ought not to tell the neighbors she goes* $\neg t^o$. *First, observe that the*

[13] The second condition of $\mathbf{React}_x$ may contain a finally eliminated sequent.

arguments a and b cannot be attacked. For that reason, we can provide ADAC*-derivations for* $\neg h^f, (\neg h, \neg t) \Rightarrow^{[!]} \neg t^o$ *and* $\top^f, \neg h^c \Rightarrow^{[!]} \neg(\top, h)$*. Hence,* $\mathcal{K} \mid\!\sim^{s} \neg t^o$ *as desired. What is more, since b attacks c, we obtain the following derivation:*

$$\dfrac{c[\mathsf{i}] \qquad \dfrac{b[\mathsf{i}] \qquad \emptyset \qquad \emptyset}{b[!]}\,\textbf{\textit{Final}}_x}{c[\bot]}\,\textbf{\textit{FinalElim}}_x$$

The derived argument $\top^f, (\top, h) \Rightarrow^{[\bot]} h^o$ *tells us that the norm* $(\top, h)$ *is* strictly inapplicable *in the violation context $\mathcal{K}$, i.e., incompatible with any maximally consistent set of* $\mathcal{K} = \langle \mathcal{F}, \mathcal{N}, \mathcal{C} \rangle$*.*

The following final derivability rule $\mathbf{Fin2}_x$ is admissible in the light of $\mathbf{FinalElim}_x$:

$$\dfrac{\Gamma_1 \Rightarrow^{[\mathsf{i}]} \Delta_1 \qquad (\forall \Gamma_2 \Rightarrow \Delta_2 \in \mathsf{Def}(\Gamma_1))\ \Gamma_2 \Rightarrow^{[\bot]} \Delta_2}{\Gamma_1 \Rightarrow^{[!]} \Delta_1}\,\mathbf{Fin2}_x$$

The rule expresses that if all of an argument's attackers are finally eliminated, the argument in question is finally acceptable. In fact, the above rule may be considered as a simplification of $\mathbf{Final}_x$. To see this, consider an arbitrary application of $\mathbf{Final}_x$ which derives $s_1[!]$. Then, for each attacker $s_2[*]$ of $s_1[\mathsf{i}]$ there exists a derivable sequent $s_3[!]$ attacking $s_2[*]$. Hence, we can apply the $\mathbf{FinalElim}_x$ rule to each pair $s_2[*]$ and $s_3[!]$ to derive $s_2[\bot]$ and subsequently apply the above rule $\mathbf{Fin2}_x$ to obtain $s_1[!]$.

Last, it can be easily seen that adding the annotation $[\bot]$ and corresponding rule $\mathbf{FinalElim}_x$ do not change the desired correspondence with DAC-induced argumentation frameworks. First, observe that finally accepted arguments $s[!]$ of a coherent ADAC-derivation $\mathcal{D}$ are member of the grounded extension $\mathcal{E}$ of the argumentation framework $\mathcal{AF}(\mathcal{D})$. The grounded extension is the minimal complete extension subset of all complete extensions, and *a fortiori*, a subset of all stable extensions (Definition 7.11). Consequently, by the conflict-freeness of complete extensions, any sequent attacked by a finally accepted sequent is excluded from any stable extension. Furthermore, Corollary 7.4 tells us that for any coherent ADAC-derivation, the status of a sequent s_1 attacked by a finally accepted sequent s_2 is permanently eliminated [e] and so, an application of $\mathbf{FinalElim}_x$ would preserve this fact by changing the status of s_1 to $[\bot]$.

We leave further investigation of finally eliminated sequents for future work.

7.6 Related Work and Future Research

In this last section, we discuss the proof systems and results presented in this chapter in light of related literature. Along the way, we point out some open questions that should be addressed in future research.

Using Annotations. Annotated versions of sequent calculi have been introduced to the literature. We briefly discuss three of them. Indrzejczak (1997) proposed a generalized sequent formalism for propositional modal logics, where sequent arrows may become annotated with modal operators together with an index to indicate the modal depth of the formulae in the sequents (sequents may also lose their annotation throughout a derivation). Similarly, Došen (1985) developed sequent-style calculi for modal logics (S5 and S) where sequent arrows may be annotated with natural numbers. Furthermore, the notion of a sequent is generalized, i.e., a sequent can have sets (of sets) of sequents on the left-hand and right-hand side of a sequent arrow. Roughly, such sequents express that from sets of sequents, one can derive other sets of sequents. To illustrate, for $\Rightarrow^1$, there are only formulae on the lhs and rhs; for $\Rightarrow^2$, there are only sets of sequents with formulae on the lhs and rhs, etcetera. The main difference with the above two approaches is that in our approach, annotations serve to indicate the derivability status of the sequent, whereas the annotations in (Došen, 1985; Indrzejczak, 1997) express and preserve information about the (modal) formulae within the sequent. Both systems characterize monotonic inference relations, whereas the annotations in AC characterize nonmonotonic inferences.

Last, Bonatti and Olivetti (2002) use sequent annotations to characterize propositional nonmonotonic logics. The resulting sequent-style calculi employ sequents with three types of sequents: monotonic sequents, non-derivability sequents, and nonmonotonic sequents. The calculi characterize circumscription logic, default logic, and autoepistemic logic. The calculi are analytic. Similar to our approach, their calculi characterize both credulous and skeptical inference. Concerning annotations, the main difference with our approach is that for Bonatti and Olivetti (2002), once a sequent arrow changes to the nonmonotonic sequent arrow, it does not change anymore.

Formal Argumentation. Various reasoning methods exist for abstract and structured (or, more specifically, logical) argumentation frameworks. We refer to (Cerutti et al., 2017) and (Besnard et al., 2020) for two extensive surveys on this subject. Most of the approaches mentioned in these review papers are based on CSP/SAT/ASP/QBF-solvers. As such, the reasoning engine is encapsulated in the solvers and often limited to specific base logics. Our approach extends standard sequent calculi using the existing notions of sequent systems (sequents, inference rules, etcetera), augmented with primary concepts from argumentation theory for a better handling of conflicts among the sequents.

Adaptive and related logics. Some AC notions, such as final acceptability, are borrowed from the dynamic proof systems developed by Arieli and Straßer (2019) and adaptive logics by Batens (2007) and Straßer (2014). We discuss both in more detail.

The formal system presented by Arieli et al. (2022a) and those in this chapter can be seen as a continuation of the dynamic proof systems developed by Arieli and Straßer (2019). Their approach contains a proof-theoretic characterization of logical argumentation. Arguments are taken as derivable sequents, and conflicts among them are captured through sequent elimination rules. They, too, take the resulting proofs as lists (and not necessarily trees). Arieli and Straßer (2019) provide a large class of elimination rules mimicking various known attack rules from the field of argumentation (e.g., undercuts, rebuttals, undermines). They differentiate between accepted and eliminated sequents by using two types of sequents (derived and non-derived). Their method, too, is highly modular with respect to the underlying language, base logic, and elimination rules. Their proof systems may be taken as the first study of sequent-style calculi for characterizing semantic extensions of formal argumentation. The systems are shown to be nonmonotonic and paraconsistent. Furthermore, logical properties of the consequence relation, such as cautious monotonicity, are shown. We leave an investigation of such additional properties for AC to future work. The main differences with the proof systems proposed in this chapter are twofold: First, we integrate metareasoning about final acceptability in the object language of the proof system, i.e., through using annotations, whereas Arieli and Straßer (2019) define final acceptability by referring to all potential extensions of a given proof and, thus, inference is not fully incorporated on the level of the proof. Second, we additionally prove a correspondence between finally accepted sequents and the grounded semantics of the corresponding argumentation framework.

It must be noted that Arieli and Straßer (2019) demonstrate soundness and completeness between dynamic derivations and stable extension in argumentation frameworks. In this chapter, we only proved one direction of the equivalence, i.e., how an AC-derivation implies the existence of a stable extension in its corresponding argumentation framework (Proposition 7.8). We leave the other direction for future work.

Open question 7.1. *Can we provide a constructive proof of how a stable extension in a given argumentation framework can be turned into a coherent* AC*-derivation?*

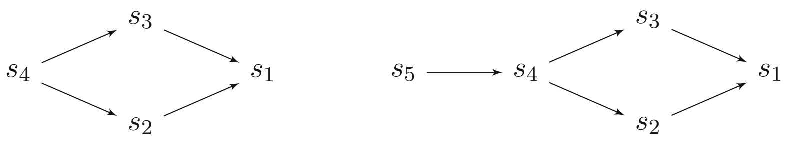
Adaptive Logic, initially developed by Batens (2007), is a formalism for defeasible reasoning. Adaptative logics are nonmonotonic. The primary mechanism concerns blocking inferences in a proof whenever the assumed knowledge base is inconsistent. Furthermore, adaptive logics were shown to correspond to Default Logic and formal argumentation (Straßer, 2014). Adaptive logics consist of a base logic, referred to as the lower limit logic, a set of abnormalities (such as $\varphi \wedge \neg\varphi$), and adaptive strategies for dealing with abnormalities. The central idea is to reason with an (inconsistent) knowledge base as "as normally as possible", which means that as few as possible abnormalities occur (Straßer, 2014). Its proof theory consists of dynamic proof systems. Proofs are lists of tuples (as for AC). The marking of a line during a proof construction means that the derived formula is defeated, i.e., inactive (for that time). In short, in a dynamic proof, one reasons as if there are no abnormalities until an abnormality is encountered. Such an encounter then yields specific lines inactive. Unmarked lines may be marked later in the derivation and vice versa. Adaptive logics have several strategies for marking lines. Like AC, dynamic proofs give a procedural way of revising the acceptability of conclusions.

A significant difference with our approach is that adaptive logics use Hilbert-style proofs. Sequent-style systems are more derivation friendly (Arieli and Straßer, 2019). Furthermore, the proofs generated by AC are *uni-directional*. This means that the status of a tuple is determined solely in view of tuples that occur further down in the proof. Adaptive logics employ a bi-directional marking where a derivation line may be marked because of a line previously occurring in the derivation. The main difference with our approach is that the machinery for updating the statuses of derived sequents in AC is included in the derivation itself and does not require an external evaluation procedure. Namely, in adaptive logics, a formula is finally derivable if, for any derivation in which the formula is marked, there is an extension of the derivation that unmarks it. In fact, by allowing inference rules to reason about the acceptability statuses of arguments, our approach fully integrates *meta-argumentative* reasoning into the object language of the proof; cf. (Jakobovits and Vermeir, 1999; Boella et al., 2009).[14]

Tree-like derivations. Our approach, as well as adaptive logics Straßer (2014) and the approach by (Arieli and Straßer, 2019), generates derivations that are not necessarily representable as tree-like structures. At the moment, AC generate collections of trees, i.e., forests. We illustrate this with an example. Consider the following coherent derivation (we leave the attacking conditions implicit since they are assumed identical to the attacking sequent):

$$\dfrac{\dfrac{\dfrac{\dfrac{s_1[\mathsf{i}] \qquad s_2[\mathsf{i}]}{s_1[\mathsf{e}]}\,\textbf{Def} \qquad \dfrac{s_2[\mathsf{i}] \qquad s_4[\mathsf{i}]}{s_2[\mathsf{e}]}\,\textbf{Def}}{s_1[\mathsf{i}]}\,\textbf{React} \qquad s_3[\mathsf{i}]}{s_1[\mathsf{e}]}\,\textbf{Def} \qquad \dfrac{s_3[\mathsf{i}] \qquad s_4[\mathsf{i}]}{s_3[\mathsf{e}]}\,\textbf{Def}}{s_1[\mathsf{i}]}\,\textbf{React}$$

We refer to this derivation as $\mathcal{D}$. The corresponding argumentation framework $\mathcal{AF}(\mathcal{D})$ is graphically represented below on the left.



$\mathcal{AF}(\mathcal{D})$ $\mathcal{AF}(\mathcal{D}')$

Now, suppose we extend the above derivation with an argument s_5 that only attacks s_4. That is, we extend $\mathcal{D}$ with the following derivation,

$$\dfrac{s_4[\mathsf{i}] \qquad s_5[\mathsf{i}]}{s_4[\mathsf{e}]}\,\textbf{Def}$$

[14] Similarly, the Deontic Argumentation Calculi, developed in Chapter 6, capture meta-argumentative aspects of defeasible (normative) reasoning concerning the applicability of norms.

resulting in a derivation $\mathcal{D}'$ (the corresponding $\mathcal{AF}(\mathcal{D}')$ is shown above on the right). The elimination of s_4 (above) triggers a revision process with $s_4 \in \mathsf{RevSeq}$. Since s_4 attacks both s_2 and s_3, we need to revise both. It can be easily seen that the resulting derivation forest $\mathcal{D}''$ cannot be turned into a tree-like structure:

$$\dfrac{\begin{matrix}\mathcal{D}\\ \vdots \\ s_1[\mathsf{i}]\end{matrix} \qquad \dfrac{\dfrac{s_2[\mathsf{i}] \qquad s_4[\mathsf{i}]}{s_2[\mathsf{e}]}\,\textbf{Def} \qquad \dfrac{s_4[\mathsf{i}] \qquad s_5[\mathsf{i}]}{s_4[\mathsf{e}]}\,\textbf{Def}}{s_2[\mathsf{i}]}\,\textbf{React}}{s_1[\mathsf{e}]}\,\textbf{React}$$

$$\dfrac{s_1[\mathsf{i}] \qquad \dfrac{\dfrac{s_3[\mathsf{i}] \qquad s_4[\mathsf{i}]}{s_3[\mathsf{e}]}\,\textbf{Def} \qquad \dfrac{s_4[\mathsf{i}] \qquad s_5[\mathsf{i}]}{s_4[\mathsf{e}]}\,\textbf{Def}}{s_3[\mathsf{i}]}\,\textbf{React}}{s_1[\mathsf{e}]}\,\textbf{Def}$$

The resulting derivation (forest) is coherent and, as desired, the accepted arguments $\{s_5, s_2, s_3\}$ form a stable extension. Notice that if an AC-derivation were restricted to tree-like structures only, the resulting derivation (the topmost derivation) would not be coherent since s_3 would have remained eliminated.

It remains an open question to determine how AC can be modified to generate tree-like proofs. One of the upshots of using trees is that the proof has a single conclusion (i.e., the root), and all applications of rules are directly related to the derivation of the conclusion.

Open question 7.2. *How can we modify* AC *and its corresponding revision process such that all resulting* AC*-derivations are tree-like?*

Proof theory for nonmonotonic deontic logics. In Section 7.5 we introduced an extension of AC incorporating the language and sequents of Deontic Argumentation Calculi from Chapter 6. The resulting Annotated Deontic Argumentation Calculi were shown to be nonmonotonic. These preliminary results open the door for an integrated approach to nonmonotonic normative reasoning (e.g., based on I/O logics). We briefly discuss other nonmonotonic proof systems for deontic logics.

Governatori and Rotolo (2006) developed—in a series of papers—defeasible deontic logic, which is a sequent-style proof system for reasoning with CTD obligations; see also (Governatori et al., 2018). Their system resolves around the notion of 'normative reparation' (cf. reparational oughts, Chapter 2, page 54). Their approach's central motivation is that norms must be violable to be meaningful. Since violations are exceptional circumstances (potentially) giving rise to new obligations, Governatori and Rotolo (2006) argue that primary obligations and their CTD obligations must be considered as generating a single norm. Roughly, if $\varphi \Rightarrow \mathcal{O}\psi$ and $\varphi, \neg\psi \Rightarrow \mathcal{O}\theta$, then this expresses the unique reparational ought $\varphi \Rightarrow \psi \otimes \theta$ which must be read as "in the context φ it is obligatory that ψ, but in case one fails to comply (violates), then it is obligatory that θ" (a $\otimes$ formula must be read from left to right). The language contains only atoms, negation, and the operator $\otimes$. Their proposed sequent-style systems modify strings of reparational oughts. Subsequently, inferences are based on extensions generated by a context and a set of reparational oughts. The logic is nonmonotonic since a change in context may ensue different obligations. We refer to the work of Governatori and Rotolo (2006) for a discussion of the differences between Defeasible Deontic Logic, Input/Output logic (Makinson and van der Torre, 2001), and the nonmonotonic deontic logic developed by Prakken and Sergot (1996). We point out that Deontic Argumentation Calculi (Chapter 6) are developed explicitly for Input/Output logics. We leave it for future work to formally investigate the relation between DAC and defeasible deontic logic.

Lellmann et al. (2021) provide a proof-theoretic approach to nonmonotonic deontic logic based on deontic theories of the ancient Sanskrit philosophy school called Mīmāṃsā (see Chapter 5 for an extensive historical introduction to Mīmāṃsā). They developed a sequent-style proof system for reasoning with obligations, prohibitions, and recommendations. Given a set of (deontic) assumptions, conflicts are dealt with via a Mīmāṃsā-inspired specificity principle. The principle is more involved than the one proposed by Horty (1997) since it references underivability statements. Moreover, it checks whether, e.g., obligations that override other obligations are themselves not overridden by again more specific conflicting obligations. Their proof system satisfies cut-elimination and is shown decidable. There is no sound and complete semantics available for this proof system. Furthermore, their calculus contains a rule for *vikalpa*, a Mīmāṃsā principle that corresponds to the nonmonotonic principle called disjunctive response (i.e., in case of conflicting commands, one is obliged to choose at least one option) (cf. Chapter 3, page 112).

Horty (1997) developed one of the first nonmonotonic accounts of normative reasoning. Although the approach does not involve proof theory. It is worth mentioning due to its close relation to Default Logic, Input/Output logic, and consequently ADAC. See the work of Parent (2011) for a correspondence result between I/O logic and Deontic Default Logic. Horty's motivation lies in commonsense normative reasoning, which often involves rules of thumb—such as "Do not harm anyone"—and thus is defeasible. The formal system developed by Horty (1997) is tailored to handling normative conflicts. It contains a deontic extension of Reiter's Default Logic (Reiter, 1980), which includes conditional oughts. Furthermore, it involves a specificity principle, i.e., an obligation is overridden if it conflicts with an obligation that has a strictly more specific antecedent. Some of the open problems posed by Horty (1997), concerning the transitivity of conditional oughts and reasoning with disjunctive contexts, were satisfactorily addressed by the constrained Input/Output formalism introduced by Makinson and van der Torre (2001).

Concerning the above, future work must be directed to the following open problem.

Open question 7.3. *What is the correspondence between maximally consistent sets of norms generated by constrained Input/Output logics (Section 6.2) and coherent derivations generated by the Annotated Deontic Argumentation Calculi* ADAC *(Section 7.5)?*

Answering the above question likewise opens the door to a formal comparison with the systems developed by Horty (1997). Positive results appear promising because our systems correspond to stable extensions in their induced argumentation frameworks (i.e., Proposition 6.2 and Proposition 7.15).

* * *

In this chapter, we introduced Annotated Calculi AC. A class of sequent-style proof systems that is highly modular with respect to its language and base logic. We showed that the consequence relation of AC is paraconsistent and nonmonotonic (Objective 1). Moreover, we demonstrated a strong correspondence with logical argumentation (Objective 2). Last, we provided promising results on extending AC to the context of defeasible normative reasoning by developing the class of nonmonotonic proof systems, called Annotated Deontic Argumentation Calculi ADAC (Objective 3). For instance, ADAC preserves the desired correspondence with semantic extensions in logical argumentation.

CHAPTER 8

Conclusion

In this final chapter, we briefly recapitulate the main results of each individual chapter. After that, we provide a more general reflection on the thesis and conclude with two promising future research directions.

8.1 Summary

This thesis is about the *logical analysis of normative reasoning*. It supports the claim that a better understanding of normative reasoning can be gained by involving agents in its formal analysis. We identified various problems in deontic logic and some novel challenges in AI. These were systematically addressed by dividing this thesis into three parts, each consisting of two chapters. We briefly summarize the main results and insights acquired in each chapter.

In Part I, we dealt with *Agency and Normative Reasoning*. We identified two general challenges. In Chapter 2, we investigated reasoning about obligation and choice in an explicitly indeterministic temporal setting (research question 1). In Chapter 3, we studied the logical relations between various readings of Ought implies Can and determined the consequences of these readings for formal normative reasoning (research question 2).

Chapter 2. We provided a sound and complete Temporal Deontic STIT logic, filling a long-standing gap in the STIT literature. We showed how the proposed relational semantics of our logic can be truth-preservingly transformed into the traditional utilitarian STIT semantics of dominance ought. We applied the logic to assess certain arguments made by Horty (2001) concerning the impact of temporal reasoning on obligations based on utility assignments. For instance, we proved that two-valued assignments are incomplete for Temporal Deontic STIT logic. Conceptually, this result shows that deliberative agency and contrary-to-duty reasoning are incompatible with two-valued utilitarian approaches in indeterministic time.

Chapter 3. We provided a comprehensive philosophical and logical investigation of *Ought implies Can* (OiC). We developed a class of sound and complete deontic STIT logics axiomatizing ten OiC interpretations. The logics were employed to provide a formal taxonomy of OiC, determining the (in)dependencies between the various interpretations. We extended the resulting logics with other metaethical principles, determining their relation to OiC. We argued that the possibility of deontic dilemmas is logically independent of OiC and showed that by adopting the principle of Deontic Contingency, various readings of OiC become equivalent, leading to strictly fewer interpretations. Last, we demonstrated how to restore some of the inferential power lost by adopting a non-normal modal approach to deontic STIT logic. Conceptually, this chapter emphasizes the importance of concepts such as ‘ability’, ‘violability’, and ‘control’ for the analysis of normative reasoning.

In Part II, we addressed *Action and Normative Reasoning*. We formulated two main challenges. In Chapter 4, we investigated the formal representation of obligations and prohibitions about instruments and studied their logical properties (research question 3). In Chapter 5, we provided an application of deontic action logic to ancient Sanskrit philosophy by analyzing Maṇḍana’s theory of deontic reasoning (research question 4).

Chapter 4. We addressed the question of instrumentality statements in the context of normative reasoning. We provided a sound and complete modal logic of action and norms to reason about such statements. In particular, we formalized and analyzed a novel yet ubiquitous norm category called *norms of instrumentality* and investigated the logical relations between this category and the well-known categories of ought-to-be and ought-to-do. In particular, we argued that the three categories are reciprocally irreducible. Furthermore, we analyzed how instrumentality statements in norms relate to metaethical principles such as No Vacuous Commands and Ought implies Can. Last, we discussed possible extensions of the resulting logic containing more refined instrumentality statements. Means-end reasoning is an outstanding feature of practical reasoning, and if norms are to influence deliberation, a proper formal understanding of instrumentality relations in the context of norms is essential. This chapter provided such an analysis.

Chapter 5. We formally analyzed the deontic theory of the acclaimed Mīmāṃsā author Maṇḍana, whose theory consists of reducing all Vedic commands to statements about actions as instruments leading to specific results. We provided a sound and complete logic capturing this deontic reduction. We showed that some general Mīmāṃsā principles are not valid in the resulting logic, and argued that this is in accordance with Maṇḍana’s view on instrumentality. We gave a logical analysis of Maṇḍana’s solution to the Śyena controversy, argued that the controversy is akin to the contemporary Gentle Murder Paradox, and satisfactorily evaluated the proposed logic on some well-known contemporary deontic puzzles. Conceptually, the results obtained in this chapter demonstrate how ancient sources can provide substantial input for developments in deontic logic.

In Part III, we took on a novel research topic concerning *Argumentation and Normative Reasoning*. We addressed several challenges. In Chapter 6, we developed a modular class of proof systems yielding argumentative characterizations of a large class of nonmonotonic Input/Output logics, and showed how the resulting formalism accommodates deontic explanations (research questions 5 and 6). In Chapter 7, we addressed the more general challenge of integrating status revision considerations of defeasible reasoning into the object level of sequent-style proof systems to yield nonmonotonic proof systems (research question 7).

Chapter 6. We laid the formal foundations for *deontic explanations* in Formal Argumentation and AI. We defined these explanations as answers to why questions such as “Why am I obliged to do X, despite my conflicting obligation to do Y?”. We introduced a highly modular proof theoretic formalism called Deontic Argumentation Calculi, which explicitly and transparently formalizes reasons, integrating meta-reasoning about the inapplicability of norms into the object language. We proved that the calculi are sound and complete with respect to a large class of monotonic Input/Output logics and demonstrated that argumentation frameworks instantiated with arguments generated by these calculi are sound and complete for the class of nonmonotonic constrained Input/Output logics. We discussed the explanatory nature of the resulting argumentation frameworks and extended the calculi with relevance rules that exclude deontic arguments containing irrelevant reasons. Conceptually, this chapter shows that the characterization of defeasible deontic reasoning using methods from formal argumentation is a promising research direction for explainability in the context of (AI) agents.

Chapter 7. We introduced Annotated Calculi, a class of proof systems highly modular with respect to their underlying base logic. Annotated Calculi incorporate revision procedures of defeasible reasoning by annotating sequents with their status and employing annotation revision rules. We showed that the consequence relations of these calculi are nonmonotonic and paraconsistent. We demonstrated correspondence between the different annotations on derivable sequents and various kinds of semantic extensions employed in formal argumentation. We extended the calculi to the context of defeasible normative reasoning, resulting in a class of nonmonotonic Annotated Deontic Argumentation Calculi. In particular, the extended calculi employ the transparent formalism of Chapter 6 facilitating deontic explanations. We provided some promising results concerning this extension. For instance, the calculi preserve the desired correspondence with formal argumentation.

8.2 General Reflection

Just as Makinson (1999) reminded us that there is “no logic of norms without attention to the normative system in which they occur” (p.32), we say that there is no accurate analysis of normative reasoning without considering the agents to which norms apply.

The thesis supports this view. In particular, the various studies conducted in this work demonstrate that central challenges of normative reasoning are ultimately related to *deliberation*, i.e., the act of practical reasoning, weighing choices, and making decisions. Let us briefly reflect on this observation.

First, Ought implies Can ensures that norms are not overdemanding for agents (Chapter 3). It requires obligations to be consistent with that which an agent can do. Thus, the principle ensures that norms *can be taken into account by the agent* when deliberating, choosing, and acting. This role of Ought implies Can is similar to that of the Mīmāṃsā principle requiring norms to be meaningful, i.e., observable and violable by agents (Chapter 5). As shown in Chapters 3–5, other metaethical principles—such as deontic contingency, no vacuous commands, and no deontic dilemmas—fulfill a similar role with respect to deliberation.

Second, defeasible mechanisms for reasoning about contrary-to-duty scenarios and deontic dilemmas are about resolving and avoiding conflicts, keeping the actual obligations implied by a normative code (jointly) consistent (Chapters 6 and 7). Thus, those mechanisms keep duties meaningful for deliberating agents bound by the code. A normative code projects an ideal image onto the world, and those who can actively shape the world according to this image are agents. Agents are, so to say, the mediators between the ideal world and the actual world (including subideal worlds). For this reason, resolving conflicts is fundamental to deliberating agents actively shaping this world.

Last, deontic explanations serve deliberation. They elucidate why the normative code requires the agent to behave in a particular way (Chapter 6). An improved understanding of how normative codes yield obligations helps the agent to understand the normative system in question better. More importantly, it helps the agent to make better-informed choices. Such explanations not only motivate compliance but also provide reasons for potential disagreement, facilitating discussion and group deliberation.

8.3 Future Research

In closing, we briefly reflect on the two most promising future research directions.

The first topic relates to the involvement of agency in defeasible normative reasoning. Most, if not all, nonmonotonic accounts of deontic logic do not include agents or actions in the analysis. This is surprising because practical reasoning, planning, and decision-making are highly defeasible, strongly depending on incomplete information, abductive statements, and rules of thumb. The nonmonotonic Input/Output logics treated in part III of this thesis also do not explicitly involve agents in the formal language.

A promising way to introduce agency to defeasible normative reasoning is by furthering our formal analysis of Ought implies Can. Agents often find themselves in situations with incomplete information about what they can and cannot do. For example, I might believe I am able to attend my band rehearsal on time, not realizing that my bike has been stolen. Once I find out that my bike is stolen, there may be consequences for the obligations that apply to me (for instance, I may be permitted to arrive late). By taking into account what agents can and cannot do when inferring obligations, we may interpret Ought implies Can as a defeasible principle. The extension of Deontic Argumentation Calculi (Chapter 6) to include an agent-sensitive language (such as the language of STIT) in reasoning defeasibly about agents and their obligations is, therefore, highly promising. An agent's abilities can then be taken as constraints blocking certain deontic inferences. As a possible application, such a formalism enables us to analyze why obligations that hold over time cease to hold due to an agent's changing abilities. The differentiation between Ought implies Can and contrary-to-duty reasoning is of particular interest in this respect. For instance, my inability to attend my band rehearsal on time due to my bike being stolen may qualify as a case of Ought implies Can defeasibility. In contrast, my inability to be on time due to (deliberately) sleeping in may be a case of violation, inducing a contrary-to-duty situation. Differentiating between inability and violation plays an essential role in studies of responsibility.

The second topic concerns deontic explanations. The investigation of explanations in the context of AI, such as formal argumentation, is relatively new. Its importance increases by the day, especially in light of autonomous AI that must reason and comply with various normative codes. We argued that adequate normative explanations require reference to aspects of normative reasoning often reserved for the meta-analysis of formal models. In Chapter 6 and 7, we pursued this line of research: we internalized reference to the applicability and inapplicability of norms, enabled reasoning about attacks and defeats between arguments, and formalized how this influences the arguments' respective acceptability status. Furthermore, we proposed ways to reason about the relevance of reasons within the language of the respective proof systems. These results are promising for further work on explanatory reasoning, internalizing other aspects of normative reasoning usually reserved for meta-analysis. In particular, future work must be directed to expanding Deontic Argumentation Calculi to reason about sets of norms, priority orderings over norms, the interplay between constitutive and regulative norms, and values promoted by norms.

What is more, the fact that explanations typically occur in the context of dialogues motivates the extension of our work to formal dialogue models. Such models employ a rich language of speech acts, including various why-questions and critical questions. Addressing deontic explanations in the setting of dialogues will enable us to model the interactive exchange of reasons, questions, and explanatory arguments, thus, tailoring the resulting explanations to the background of the explainee.

Index

List of Symbols

Modalities

$[i]$	Def. 2.1, p. 29
$\langle i \rangle$	p. 29
$[i]^d$	p. 27
$[Ag]$	Def. 2.1, p. 29
$\Box$	Def. 2.1, p. 29
$\Diamond$	p. 29
G	Def. 2.1, p. 29
H	Def. 2.1, p. 29
F	p. 29
P	p. 29
$\otimes_i$	Def. 2.1, p. 29 and Def. 3.1, p. 85
$\otimes_i^d$	p. 65
$\boxed{\mathsf{S}}$	Def. 4.2, p. 132
$\boxed{\mathsf{A}}$	Def. 4.2, p. 132
$\Diamond\mkern-11mu{\scriptstyle\mathsf{S}}$	p. 132
$\Diamond\mkern-11mu{\scriptstyle\mathsf{A}}$	p. 132
$[\Delta_i]$	p. 133
$[\Delta_i]^{would}$	p. 133
$[\Delta_i]^{could}$	p. 134
$[\Delta_i]^{will}$	p. 134
$\mathcal{F}_i(\varphi)$	p. 146
$\mathcal{O}_i(\varphi)$	p. 146
$\mathcal{F}_i[\Delta]$	p. 147
$\mathcal{O}_i[\Delta]$	p. 147
$\mathcal{F}_i[\Delta](\varphi)$	p. 149
$\mathcal{O}_i[\Delta](\varphi)$	p. 148
$\blacksquare$	p. 158
$\blacklozenge$	p. 158
$[\Delta_i]_n^{p-instr^*}$	p. 160
$[\Delta_i]_n^{exc-instr^*}$	p. 160
$\boxed{\mathsf{U}}$	Def. 5.2, p. 178
$\Diamond\mkern-11mu{\scriptstyle\mathsf{U}}$	p. 178
$\mathcal{I}(\Delta/\varphi/\chi)$	Def. 5.10, p. 187
$\mathcal{O}(\Delta/\chi)$	Def. 5.10, p. 187
$\mathcal{F}(\Delta/\chi)$	Def. 5.10, p. 187
$\mathcal{E}(\Delta/\varphi/\chi)$	Def. 5.10, p. 187

Other Syntax

$p, q, r, \ldots$	p. 29
$\varphi, \psi, \gamma, \ldots$	p. 29
$\Delta, \Gamma, \Sigma, \ldots$	p. 29
$\neg, \wedge, \vee, \rightarrow, \equiv$	p. 29
$\top, \bot$	p. 29
$-, \cup, \cap, \setminus, =, \in, \subseteq$	set-theoretic syntax
$w, v, u, \ldots$	Def. 2.4, p. 33
$X, Y, Z, \ldots$	Def. 3.4, p. 89
$\delta_i, \Delta \cup \Gamma, \overline{\Delta}, \Delta \& \Gamma$	Def. 4.1, page 132
$\mathtt{v}_i$	Def. 4.2, p. 132
$\mathtt{d}_i^\delta$	Def. 4.2, p. 132
$\mathtt{e}_i$	p. 158
$\mathtt{P}$	Def. 5.2, p. 178
$\mathtt{R}$	Def. 5.2, p. 178
$\varphi^f, \varphi^o, \varphi^c$	Def. 6.1, p. 216
(φ, ψ)	Def. 6.1, p. 216
$\neg(\varphi, \psi)$	Def. 6.6, p. 224

$\mid\!\sim^{s}_{\mathcal{R},\mathsf{L}}, \mid\!\sim^{c}_{\mathcal{R},\mathsf{L}}$	Def. 6.5, p. 223
$\vdash_{\mathsf{LC}}$	Def. 6.7, p. 225
$\vdash_{\mathcal{S}}$	Def. 6.8, p. 225
$\mathcal{AF} \mid\!\sim^{s}, \mathcal{AF} \mid\!\sim^{s^*},$ $\mathcal{AF} \mid\!\sim^{c}$	Def. 6.10, p. 241
$\mid\!\sim^{s}_{\mathsf{AC}}$	Def. 7.9, p. 265
$\mid\!\sim^{s}_{\mathsf{ADAC}_{\mathcal{S}}}$	Def. 7.15, p. 282

Modal semantics

$\mathcal{R}_{\Box}$	Def. 2.4, p. 33
$\mathcal{R}_{[i]}$	Def. 2.4, p. 33
$\mathcal{R}_{\otimes_i}$	Def. 2.4, p. 33
$\mathcal{R}_{[Ag]}$	Def. 2.4, p. 33
$\mathcal{R}_{\mathsf{G}}$	Def. 2.4, p. 33
$\mathcal{R}_{\mathsf{H}}$	Def. 2.4, p. 33
$\mathcal{R}^{s}_{[i]}$	p. 55
W	Def. 2.4, p. 33
V	Def. 2.4, p. 33
$\mathcal{N}_{\otimes_i}$	Def. 3.4, p. 89
$W_{t(\Delta)}$	Def. 4.5, p. 137
$W_{\mathtt{d}^{\delta}_{i}}, W_{\mathtt{v}_i}$	Def. 4.6, p. 137
$\mathcal{R}_{\boxed{\mathsf{S}}}$	Def. 4.6, p. 137
$\mathcal{R}_{\boxed{\mathsf{A}}}$	Def. 4.6, p. 137
$W_{\mathtt{P}}$	Def. 5.5, p. 180
$W_{\mathtt{R}}$	Def. 5.5, p. 180
$\mathcal{R}_{\boxed{\mathsf{U}}}$	Def. 5.5, p. 180

Argumentation

$\mathcal{AF}$	Def. 6.9, p. 240
$\mathcal{AF}_{\mathcal{S}}(\mathcal{K})$	Def. 6.9, p. 240
Arg	Def. 6.9, p. 240
Att	Def. 6.9, p. 240
$\mathcal{E}, \mathcal{E}_a, \mathcal{E}^+, \mathcal{E}^-$	p. 242
$\mathsf{Derived}, \mathsf{Accept},$ Final	Def. 7.10, p. 274

Other symbols

Atoms	Def. 2.1, p. 29
Agents	Def. 2.1, p. 29
$\preceq, \prec$	p. 56
$\leq, <$	p. 56
util	Def. 2.13, p. 55
util^{m}	p. 64
util^{h}	p. 64
h	p. 64
$\mathsf{M}, \mathsf{C}, \mathsf{N}$	p. 88
$\|\varphi\|_{\mathfrak{M}}$	Def. 3.5, page 90
$\{\!\|\varphi\|\!\}_{\mathfrak{M}}$	Def. 3.8, page 95
Act	Def. 4.1, page 132
Vio	p. 132
Wit_i	p. 132
t	Def. 4.3, p. 133
maxfam	Def. 6.5, p. 223
$\mathsf{Connectives}$	Def. 7.1, p. 258
$\mathcal{D}$	Def. 7.5, p. 263
T	Def. 7.5, p. 263
RevSeq	Def. 7.7, p. 264

Acronyms

AC	Annotated Calculi, p. 256
ADAC	Annotated Deontic Argumentation Calculi, p. 279
$\mathcal{AF}$	Argumentation Framework, p. 240
AI	Artificial Intelligence, p. 2
ASTIT	Achievement STIT, p. 71
BDI	Belief-Desire-Intention, p. 6
BT+AC	Branching Time with Agential Choice, p. 23
CTD	Contrary-To-Duty, p. 14
DAC	Deontic Argumentation Calculi, p. 211
DCg	Deontic Contingency, p. 102
DCs	Deontic Consistency, p. 102
DS$_n$	Deontic STIT Logic, p. 26
IoA	Independence of Agents, p. 29
I/O	Input/Output, p. 9
KR	Knowledge Representation and Reasoning, p. 12
LAN	Logic of Action and Norms, p. 124
LM	Logic of Maṇḍana, p. 170
MCS	Maximally Consistent Set, p. 42
XSTIT	Next STIT, p. 71
NCbUH	No Choice between Undivided Histories, p. 29
NDD	No Deontic Dilemmas, p. 102

NVC	No Vacuous Commands, p. 102
NorMAS	Normative Multi-agent Systems, p. 3
OiA	Ought implies Ability, p. 81
OiA+O	Ought implies Ability and Opportunity, p. 83
OiC	Ought implies Can, p. 75
OiCtrl	Ought implies Control, p. 83
OiLP	Ought implies Logical Possibility, p. 79
OiNA	Ought implies Normatively Able, p. 84
OiNC	Ought implies Normatively Can, p. 84
OiO	Ought implies Opportunity, p. 82
OiRz	Ought implies Realizability, p. 80
OiRef	Ought implies Refrainability, p. 82
OiV	Ought implies Violability, p. 81
OS$_n$	Logic of Ought-implies-Can, p. 77
PDeL	Deontic Propositional Dynamic Logic, p. 124
PDL	Propositional Dynamic Logic, p. 6
PMS	Pūrva Mīmāṃsā Sūtra, p. 172
ŚBh	Śabara's *Bhāṣya*, p. 172
SDL	Standard Deontic Logic, p. 13
STIT	Seeing To It That, p. 23
TDS$_n$	Temporal Deontic STIT Logic, p. 23
TUS$_n$	Temporal Utilitarian STIT Logic, p. 54
US$_n$	Utilitarian STIT Logic, p. 54

Bibliography

Abarca, Aldo Iván Ramírez and Jan Broersen (2019). "A logic of objective and subjective oughts". In: *Logics in Artificial Intelligence - 16th European Conference, JELIA 2019*. Ed. by Francesco Calimeri, Nicola Leone, and Marco Manna. Vol. 11468. Lecture Notes in Computer Science. Springer, pp. 629–641.

Abe, Jair Minoro, Kazumi Nakamatsu, and João Inácio da Silva Filho (2019). "Three decades of paraconsistent annotated logics: a review paper on some applications". In: *Proceedings of the 23rd International Conference Knowledge-Based and Intelligent Information & Engineering Systems (KES-2019)*. Vol. 159. Procedia Computer Science. Elsevier, pp. 1175–1181.

Abraham, Michael, Dov Gabbay, and Uri Schild (2011). "Obligations and prohibitions in Talmudic deontic logic". In: *Artificial Intelligence and Law* 19 (2-3), pp. 117–148. DOI: 10.1007/s10506-011-9109-0.

van Ackeren, Marcel and Michael Kühler (2015). "Ethics on (the) Edge? Introduction to Moral Demandingness and 'Ought Implies Can'". In: *The Limits of Moral Obligation*. Ed. by Marcel van Ackeren and Michael Kühler. Routledge, pp. 1–18.

Alchourrón, Carlos E. (1996). "Detachment and defeasibility in deontic logic". In: *Studia Logica* 57.1, pp. 5–18.

Alchourrón, Carlos E., Peter Gärdenfors, and David Makinson (1985). "On the logic of theory change: Partial meet contraction and revision functions". In: *The journal of symbolic logic* 50.2, pp. 510–530.

Anderson, Alan Ross (1958). "A reduction of deontic logic to alethic modal logic". In: *Mind* 67.265, pp. 100–103.

Anderson, Alan Ross and Omar Khayyam Moore (1957). "The formal analysis of normative concepts". In: *American Sociological Review* 22.1, pp. 9–17. DOI: 10.2307/2088759.

Anglberger, Albert J.J. (2008). “Dynamic deontic logic and its paradoxes”. In: *Studia Logica* 89.3, pp. 427–435.

Åqvist, Lennart (1969). “Improved formulations of act-utilitarianism”. In: *Noûs* 3.3, pp. 299–323.

– (1974). “A new approach to the logical theory of actions and causality”. In: *Logical theory and semantic analysis*. Springer, pp. 73–91.

– (2002). “Old foundations for the logic of agency and action”. In: *Studia Logica* 72.3, pp. 313–338.

Arieli, Ofer, Kees van Berkel, and Christian Straßer (2022a). “Annotated Sequent Calculi for Paraconsistent Reasoning and Their Relations to Logical Argumentation”. In: *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*. Ed. by Lud De Raedt. International Joint Conferences on Artificial Intelligence Organization, pp. 2532–2538. DOI: `10.24963/ijcai.2022/351`.

Arieli, Ofer, AnneMarie Borg, Matthis Hesse, and Christian Straßer (2022b). “Explainable Logic-Based Argumentation”. In: *Frontiers in Artificial Intelligence and Applications: Computational Models of Argument, proceedings (COMMA22)*. Ed. by Francesca Toni, Sylwia Polberg, Richard Booth, Martin Caminada, and Hiroyuki Kido. Vol. 353. IOS press, pp. 32 –43. DOI: `10.3233/FAIA220139`.

Arieli, Ofer, AnneMarie Borg, Jesse Heyninck, and Christian Straßer (2021). “Logic-Based Approaches to Formal Argumentation”. In: *Handbook of Formal Argumentation, Volume 2*. Ed. by Dov Gabbay, Massimiliano Giacomin, Guillermo R. Simari, and Matthias Thimm. College Publications, pp. 1793–1898.

Arieli, Ofer and Christian Straßer (2015). “Sequent-based logical argumentation”. In: *Argument and Computation* 6.1, pp. 73–99.

– (2019). “Logical argumentation by dynamic proof systems”. In: *Theoretical Computer Science* 781, pp. 63–91. DOI: `https://doi.org/10.1016/j.tcs.2019.02.019`.

Arioua, Abdallah and Madalina Croitoru (2015). “Formalizing explanatory dialogues”. In: *International Conference on Scalable Uncertainty Management*. Ed. by Christoph Beierle and Alex Dekhtyar. Springer, pp. 282–297.

Aristotle (2000). *Aristotle: Nicomachean Ethics*. Ed. by Karl Ameriks and Desmond Clarke. Cambridge University Press.

Arkoudas, Konstantine, Selmer Bringsjord, and Paul Bello (2005). “Toward ethical robots via mechanized deontic logic”. In: *AAAI fall symposium on machine ethics*. The AAAI Press, Menlo Park, pp. 17–23.

Armgardt, Matthias, Emiliano Lorini, and Giovanni Sartor (2018). “Reasoning about conditions in STIT logic”. In: *14th International Conference on Deontic Logic and Normative Systems (DEON 2018)*. Ed. by Jan Broersen, Cleo Condoravdi, Shyam Nair, and Gabriella Pigozzi. College Publications, pp. 15–32.

Audi, Robert (1989). *Practical reasoning*. Routledge.

Bach, Kent (2010). “Refraining, omitting, and negative acts”. In: *A Companion to the Philosophy of Action*. Ed. by Timothy O’Connor and Constantine Sandis, pp. 50–57.

Balbiani, Philippe, Andreas Herzig, and Nicolas Troquard (2008). “Alternative axiomatics and complexity of deliberative STIT theories”. In: *Journal of Philosophical Logic* 37.4, pp. 387–406.

Baroni, Pietro, Martin Caminada, and Massimiliano Giacomin (2011). “An introduction to argumentation semantics”. In: *The knowledge engineering review* 26.4, pp. 365–410.

Baroni, Pietro, Dov Gabbay, Massimiliano Giacomin, and Leendert van der Torre (2018). *Handbook of Formal Argumentation, Volume 1*. United Kingdom: College Publications.

Bartha, Paul (1993). “Conditional obligation, deontic paradoxes, and the logic of agency”. In: *Annals of Mathematics and Artificial Intelligence* 9, pp. 1–23. DOI: 10.1007/BF01531259.

Batens, Diderik (2007). “A universal logic approach to adaptive logics”. In: *Logica universalis* 1.1, pp. 221–242.

Beirlaen, Mathieu, Christian Straßer, and Jesse Heyninck (2018). “Structured argumentation with prioritized conditional obligations and permissions”. In: *Journal of Logic and Computation* 29.2, pp. 187–214.

Belnap, Nuel (1991). “Backwards and forwards in the modal logic of agency”. In: *Philosophy and phenomenological research* 51.4, pp. 777–807.

Belnap, Nuel and Mitchell Green (1994). “Indeterminism and the thin red line”. In: *Philosophical perspectives* 8, pp. 365–388.

Belnap, Nuel and Michael Perloff (1988). “Seeing to it that: a canonical form for agentives”. In: *Theoria* 54.3, pp. 175–199.

Belnap, Nuel, Michael Perloff, and Ming Xu (2001). *Facing the future: agents and choices in our indeterminist world.* Oxford University Press, Oxford.

Bench-Capon, Trevor J.M. (2002). "Value-based argumentation frameworks". In: *In Proceedings of Non Monotonic Reasoning.* Ed. by Salem Benferhat and Enrico Giunchiglia, pp. 444–453.

– (2003). "Persuasion in practical argument using value-based argumentation frameworks". In: *Journal of Logic and Computation* 13.3, pp. 429–448.

Bench-Capon, Trevor J.M. and Giovanni Sartor (2003). "A model of legal reasoning with cases incorporating theories and values". In: *Artificial Intelligence* 150.1-2, pp. 97–143.

Bentham, Jeremy (1996). *The collected works of Jeremy Bentham: An introduction to the principles of morals and legislation (1789).* Ed. by James Henderson Burns and Herbert L. A. Hart. Clarendon Press.

van Berkel, Kees, Agata Ciabattoni, Elisa Freschi, Francesca Gulisano, and Maya Olszewski (2021a). "The Gentle Murder Paradox in Sanskrit Philosophy". In: *Deontic Logic and Normative Systems - 15th International Conference, (DEON 2020/21).* Ed. by Fenrong Liu, Alessandra Marra, Paul Portner, and Frederik Van De Putte. College publications, pp. 17–35.

– (2022a). "Deontic paradoxes in Mīmāṃsā logics: there and back again". In: *Journal of Logic, Language, and Information.* DOI: https://doi.org/10.1007/s10849-022-09375-w.

van Berkel, Kees, Agata Ciabattoni, Elisa Freschi, and Sanjay Modgil (2019). "Evaluating Networks of Arguments: A Case Study in Mīmāṃsā Dialectics". In: *Logic, Rationality, and Interaction - 7th International Workshop, LORI 2019.* Ed. by Patrick Blackburn, Emiliano Lorini, and Meiyun Guo. Vol. 11813. Lecture Notes in Computer Science. Springer, pp. 355–369. DOI: 10.1007/978-3-662-60292-8_26.

van Berkel, Kees, Dov Gabbay, and Leendert van der Torre (2021b). "If You Want to Smoke, Don't Buy Cigarettes: Near-Anankastics, Contexts, and Hyper Modality". In: *Deontic Logic and Normative Systems - 15th International Conference, (DEON 2020/21).* Ed. by Fenrong Liu, Alessandra Marra, Paul Portner, and Frederik Van De Putte. College publications, pp. 36–55.

van Berkel, Kees and Tim Lyon (2019a). "A Neutral Temporal Deontic STIT Logic". In: *Logic, Rationality, and Interaction - 7th International Workshop, LORI 2019.* Ed. by Patrick Blackburn, Emiliano Lorini, and Meiyun Guo. Vol. 11813. Lecture Notes in

Computer Science. Springer, pp. 340–354. DOI: 10.1007/978-3-662-60292-8_25.

– (2019b). "Cut-Free Calculi and Relational Semantics for Temporal STIT Logics". In: *Logics in Artificial Intelligence - 16th European Conference, JELIA 2019*. Ed. by Francesco Calimeri, Nicola Leone, and Marco Manna. Vol. 11468. Lecture Notes in Computer Science. Springer, pp. 803–819. DOI: 10.1007/978-3-030-19570-0_52.

– (2021). "The Varieties of Ought-Implies-Can and Deontic STIT Logic". In: *Deontic Logic and Normative Systems - 15th International Conference, (DEON 2021/21)*. Ed. by Fenrong Liu, Alessandra Marra, Paul Portner, and Frederik Van De Putte. College publications, pp. 56–76.

van Berkel, Kees, Tim Lyon, and Francesco Olivieri (2020). "A Decidable Multi-agent Logic for Reasoning About Actions, Instruments, and Norms". In: *Logic and Argumentation - Third International Conference, (CLAR 2020)*. Ed. by Mehdi Dastani, Huimin Dong, and Leendert van der Torre. Vol. 12061. Lecture Notes in Computer Science. Springer, pp. 219–241. DOI: 10.1007/978-3-030-44638-3_14.

van Berkel, Kees, Tim Lyon, and Matteo Pascucci (2022b). *A logical analysis of instrumentality judgments: means-end relations in the context of experience and expectations*. arXiv. DOI: 10.48550/ARXIV.2209.02287.

van Berkel, Kees and Matteo Pascucci (2018). "Notions of instrumentality in agency logic". In: *International Conference on Principles and Practice of Multi-Agent Systems (PRIMA 2018)*. Ed. by Tim Miller, Nir Oren, Yuko Sakurai, Itsuki Noda, Bastin Tony Roy Savarimuthu, and Tran Cao Son. Springer, pp. 403–419. DOI: 10.1007/978-3-030-03098-8_25.

van Berkel, Kees and Christian Straßer (2022). "Reasoning With and About Norms in Logical Argumentation". In: *Frontiers in Artificial Intelligence and Applications: Computational Models of Argument, proceedings (COMMA22)*. Ed. by Francesca Toni, Sylwia Polberg, Richard Booth, Martin Caminada, and Hiroyuki Kido. Vol. 353. IOS press, pp. 332 –343. DOI: 10.3233/FAIA220164.

Besnard, Philippe, Claudette Cayrol, and Marie-Christine Lagasquie-Schiex (2020). "Logical theories and abstract argumentation: A survey of existing works". In: *Journal of Argument and Computation* 11.1-2, pp. 41–102.

Bex, Floris and Katarzyna Budzynska (2012). "Argumentation and explanation in the context of dialogue". In: *Explanation-aware Computing, Proceedings of the Seventh*

International ExaCt workshop. Ed. by Thomas Roth-Berghofer, David B. Leake, and Jörg Cassens. Vol. 9, pp. 6–10.

Bex, Floris and Douglas Walton (2016). "Combining explanation and argumentation in dialogue". In: *Argument & Computation* 7.1, pp. 55–68.

Black, Elizabeth and Anthony Hunter (2007). "A generative inquiry dialogue system". In: *Proceedings of the 6th international joint conference on Autonomous agents and multiagent systems (AAMAS 2007)*, pp. 1–8. DOI: 10.1145/1329125.1329417.

Blackburn, Patrick, Maarten de Rijke, and Yde Venema (2004). *Modal logic*. Cambridge tracts in theoretical computer science 53. Cambridge University Press. DOI: 10.1017/CBO9781107050884.

Bobrow, Daniel G., ed. (1980). *Special Issue on Non-Monotonic Logic*. Vol. 13. Artificial Intelligence Journal.

Bochman, Alexander (2014). "Dynamic causal calculus". In: *Fourteenth International Conference on the Principles of Knowledge Representation and Reasoning (KR 2014)*. Ed. by Thomas Eiter Chitta Baral Giuseppe De Giacomo. AAAI Publications, pp. 188–197.

– (2021). *A Logical Theory of Causality*. MIT Press.

Boella, Guido, Dov Gabbay, Leendert van der Torre, and Serena Villata (2009). "Meta-Argumentation Modelling I: Methodology and Techniques". In: *Studia Logica* 93.2–3, 297–355. DOI: 10.1007/s11225-009-9213-2.

Boella, Guido, Leendert van der Torre, and Harko Verhagen (2008). "Introduction to the special issue on normative multiagent systems". In: *Autonomous Agents and Multi-Agent Systems* 17.1, pp. 1–10.

Bonatti, Piero Andrea and Nicola Olivetti (2002). "Sequent calculi for propositional nonmonotonic logics". In: *ACM Transactions on Computational Logic (TOCL)* 3.2, pp. 226–278.

Borg, AnneMarie and Floris Bex (2021). "A Basic Framework for Explanations in Argumentation". In: *IEEE Intelligent Systems*, pp. 25–35. DOI: 10.1109/MIS.2021.3053102.

Bratman, Michael (1981). "Intention and means-end reasoning". In: *The Philosophical Review* 90.2, pp. 252–265.

Braüner, Torben (2022). “Hybrid Logic”. In: *The Stanford Encyclopedia of Philosophy*. Ed. by Edward N. Zalta. Spring 2022. Metaphysics Research Lab, Stanford University.

Broersen, Jan (2003). “Modal action logics for reasoning about reactive systems”. PhD thesis. Vrije Universiteit Amsterdam.

– (2004). “Action negation and alternative reductions for dynamic deontic logics”. In: *Journal of applied logic* 2.1, pp. 153–168.

– (2008). “A logical analysis of the interaction between ‘obligation-to-do’ and ‘knowingly doing’”. In: *International Conference on Deontic Logic in Computer Science*. Springer, pp. 140–154.

– (2011a). “Deontic epistemic stit logic distinguishing modes of mens rea”. In: *Journal of Applied Logic* 9.2, pp. 137–152.

– (2011b). “Making a start with the stit logic analysis of intentional action”. In: *Journal of philosophical logic* 40.4, pp. 499–530.

– (2014). “On the reconciliation of logics of agency and logics of event types”. In: *Krister Segerberg on Logic of Actions*. Ed. by Robert Trypuz. Springer, pp. 41–59.

Broersen, Jan, Mehdi Dastani, Joris Hulstijn, Zisheng Huang, and Leendert van der Torre (2001). “The BOID architecture: conflicts between beliefs, obligations, intentions and desires”. In: *Proceedings of the fifth international conference on Autonomous agents*. Ed. by Elisabeth André, Sandip Sen, Claude Frasson, and Jörg P. Müller, pp. 9–16.

Broersen, Jan, Mehdi Dastani, Joris Hulstijn, and Leendert van der Torre (2002). “Goal generation in the BOID architecture”. In: *Cognitive Science Quarterly* 2.3-4, pp. 428–447.

Broersen, Jan, Frank Dignum, Virginia Dignum, and John-Jules Ch. Meyer (2004). “Designing a deontic logic of deadlines”. In: *International Workshop on Deontic Logic in Computer Science*. Ed. by Alessio Lomuscio and Donald Nute. Springer, pp. 43–56.

Broersen, Jan, Dov Gabbay, Andreas Herzig, Emiliano Lorini, John-Jules Ch. Meyer, Xavier Parent, and Leendert van der Torre (2013). “Deontic logic”. In: *Agreement technologies*. Springer, pp. 171–179.

Broersen, Jan, Andreas Herzig, and Nicolas Troquard (2006). “Embedding alternating-time temporal logic in strategic logic of agency”. In: *Journal of logic and computation* 16.5, pp. 559–578.

Broersen, Jan and Leendert van der Torre (2011). "Ten problems of deontic logic and normative reasoning in computer science". In: *Lectures on Logic and Computation*. Springer, pp. 55–88.

Brown, Mark A. (1988). "On the logic of ability". In: *Journal of philosophical logic*, pp. 1–26.

Brunero, John (2018). "Reasons, Evidence, and Explanations". In: *Oxford Handbooks Online*. Ed. by Daniel Star, pp. 321–341. DOI: 10.1093/oxfordhb/9780199657889.013.15.

Caminada, Martin (2004). "For the sake of the argument: explorations into argument-based reasoning". PhD thesis. Vrije Universiteit Amsterdam.

– (2017). "Argumentation semantics as formal discussion". In: *Handbook of Formal Argumentation, Volume 1*. Ed. by Pietro Baroni, Dov Gabbay, Massimiliano Giacomin, and Leendert van der Torre. College Publications, pp. 487–518.

Caminada, Martin and Leila Amgoud (2007). "On the evaluation of argumentation formalisms". In: *Artificial Intelligence* 171.5-6, pp. 286–310.

Castañeda, Hector-Neri (1972). "On the semantics of the ought-to-do". In: *Semantics of natural language*. Ed. by Donald Davidson and Gilbert Harman. Vol. 40. Synthese Library. Springer, pp. 675–694. DOI: 10.1007/978-94-010-2557-7_21.

– (1981). "The Paradoxes of Deontic Logic: The Simplest Solution to All of Them in One Fell Swoop". In: *New Studies in Deontic Logic: Norms, Actions, and the Foundations of Ethics*. Ed. by Risto Hilpinen. Springer, pp. 37–85. DOI: 10.1007/978-94-009-8484-4_2.

Cerutti, Federico, Sarah Alice Gaggl, Matthias Thimm, and Johannes Peter Wallner (2017). "Foundations of Implementations for Formal Argumentation". In: *Journal of Applied Logics - IfCoLog Journal of Logics and their Applications* 4.8, pp. 2623–2706.

Chellas, Brian F. (1980). *Modal Logic*. Cambridge: Cambridge University Press. DOI: 10.1017/CBO9780511621192.

Chisholm, Roderick M. (1963). "Contrary-to-duty imperatives and deontic logic". In: *Analysis* 24.2, pp. 33–36. DOI: 10.1093/analys/24.2.33.

Chopra, Amit, Leendert van der Torre, and Harko Verhagen (2018). *Handbook of Normative Multiagent Systems*. United Kingdom: College Publications.

Ciabattoni, Agata, Elisa Freschi, Francesco A. Genco, and Björn Lellmann (2015). "Mīmāṃsā Deontic Logic: Proof Theory and Applications". In: *Automated Reasoning with Analytic Tableaux and Related Methods*. Ed. by Hans De Nivelle. Springer, pp. 323–338. DOI: 10.1007/978-3-319-24312-2_22.

Ciabattoni, Agata, Xavier Parent, and Giovanni Sartor (2021). "A Kelsenian Deontic Logic". In: *Frontiers in Artificial Intelligence and Applications, Legal Knowledge and Information Systems*. Ed. by Erich Schweighofer. Vol. 346. IOS Press, pp. 141–150. DOI: 10.3233/FAIA210330.

Ciuni, Roberto and Emiliano Lorini (2018). "Comparing semantics for temporal STIT logic". In: *Logique et Analyse* 61.243, pp. 299–339.

Clarke, David S. (1987). *Practical inferences*. Routledge Kegan & Paul.

Condoravdi, Cleo and Sven Lauer (2016). "Anankastic conditionals are just conditionals". In: *Semantics & Pragmatics* 9, pp. 1–61. DOI: 10.3765/sp.9.8.

Conee, Earl (1982). "Against moral dilemmas". In: *The Philosophical Review* 91.1, pp. 87–97.

Copp, David (2017). "'Ought' implies 'can', blameworthiness, and the principle of alternate possibilities". In: *Moral responsibility and alternative possibilities*. Routledge, pp. 265–299.

Craven, Robert and Marek Sergot (2008). "Agent strands in the action language nC+". In: *Journal of Applied Logic* 6.2, pp. 172–191.

Čyras, Kristijonas, Antonio Rago, Emanuele Albini, Pietro Baroni, and Francesca Toni (2021). *Argumentative XAI: A Survey*. arXiv. DOI: 10.48550/ARXIV.2105.11266.

Dahl, Norman O. (1974). "Ought implies can and deontic logic". In: *Philosophia* 4.4, pp. 485–511.

Dalmonte, Tiziano, Charles Grellois, and Nicola Olivetti (2021). "Proof systems for the logics of bringing-it-about". In: *Deontic Logic and Normative Systems - 15th International Conference, (DEON 2021/21)*. Ed. by Fenrong Liu, Alessandra Marra, Paul Portner, and Frederik Van De Putte, pp. 114–132.

d'Altan, Piero, John-Jules Ch. Meyer, and Roelf Johannes Wieringa (1996). "An integrated framework for ought-to-be and ought-to-do constraints". In: *Artificial Intelligence and Law* 4.2, pp. 77–111.

Dastani, Mehdi (2008). “2APL: a practical agent programming language”. In: *Autonomous Agents and Multi-agent Systems* 16.3, pp. 214–248.

Delgrande, James P. (2020). “A Preference-Based Approach to Defeasible Deontic Inference.” In: *Proceedings of the 17th International Conference on Principles of Knowledge Representation and Reasoning (KR 2020)*. Ed. by Diego Calvanese, Esra Erdem, and Michael Thielscher, pp. 326–335.

Dignum, Frank and Ruurd Kuiper (1997). “Combining dynamic deontic logic and temporal logic for the specification of deadlines”. In: *Proceedings of the Thirtieth Hawaii International Conference on System Sciences*. Ed. by Jr. Ralph H. Sprague. Vol. 5. IEEE, pp. 336–346.

Dong, Huimin, Beishui Liao, and Leendert van der Torre (2020). “Kratzer Style Deontic Logics in Formal Argumentation”. In: *Proceedings of the 18th International Workshop on Non-Monotonic Reasoning, Workshop Notes (NMR 2020)*. Ed. by Maria Vanina Martínez and Ivan Varzinczak, pp. 246–255.

Došen, Kosta (1985). “Sequent-Systems for Modal Logic”. In: *Journal of Symbolic Logic* 50.1, pp. 149–168.

Dung, Phan Minh (1995). “On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games”. In: *Artificial Intelligence* 77.2, pp. 321–357.

Dung, Phan Minh, Robert A. Kowalski, and Francesca Toni (2009). “Assumption-based argumentation”. In: *Argumentation in artificial intelligence*. Springer, pp. 199–218.

van Eck, Job A. (1982). “A System of temporally relative modal and deontic predicate logic and its philosophical applications”. In: *Logique et Analyse* 25.99, pp. 249–290.

Elgesem, Dag (1997). “The modal logic of agency”. In: *Nordic Journal of Philosophical Logic*, pp. 1–46.

Fan, Xiuyi and Francesca Toni (2014). “A general framework for sound assumption-based argumentation dialogues”. In: *Artificial Intelligence* 216.0, pp. 20–54. DOI: 10.1016/j.artint.2014.06.001.

– (2015a). “On computing explanations in argumentation”. In: *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence (AAAI 2015)*. Ed. by Blai Bonet and Sven Koenig, pp. 1496–1502.

– (2015b). “On explanations for non-acceptable arguments”. In: *International Workshop on Theory and Applications of Formal Argumentation (TAFA 2015)*. Ed. by Elizabeth Black, Sanjay Modgil, and Nir Oren. Springer, pp. 112–127.

Feldhus, Nils, Ajay Madhavan Ravichandran, and Sebastian Möller (2022). *Mediators: Conversational Agents Explaining NLP Model Behavior*. arXiv. DOI: 10.48550/ARXIV.2206.06029.

Fischer, Michael J. and Richard E. Ladner (1979). “Propositional dynamic logic of regular programs”. In: *Journal of computer and system sciences* 18.2, pp. 194–211. DOI: 10.1016/0022-0000(79)90046-1.

Floridi, Luciano and Jeff W. Sanders (2004). “On the morality of artificial agents”. In: *Minds and machines* 14.3, pp. 349–379.

Forrester, James William (1984). “Gentle Murder, or the adverbial Samaritan”. In: *Journal of Philosophy* 81.4, pp. 193–197. DOI: 10.2307/2026120.

Freschi, Elisa (2010). “Indian Philosophers”. In: *A Companion to the Philosophy of Action*. Ed. by Timothy O’Connor and Constantine Sandis. John Wiley & Sons, pp. 419–428.

– (2018). “The role of *paribhāṣā*s in Mīmāṃsā: rational rules of textual exegesis”. In: *Asiatische Studien/Études Asiatiques* 72.2. Ed. by Gianni Pellegrini, pp. 567–595. DOI: 10.1515/asia-2018-0018.

Freschi, Elisa, Agata Ciabattoni, Francesco A. Genco, and Björn Lellmann (2017). “Understanding Prescriptive Texts: Rules and Logic as Elaborated by the Mīmāṃsā School”. In: *Journal of World Philosophies* 2.1, pp. 47–66. DOI: 10.2979/jourworlphil.2.1.05.

Freschi, Elisa, Andrew Ollett, and Matteo Pascucci (2019). “Duty and Sacrifice: A Logical Analysis of the Mīmāṃsā Theory of Vedic Injunctions”. In: *History and Philosophy of Logic* 40.4, pp. 323–354. DOI: 10.1080/01445340.2019.1615366.

Freschi, Elisa and Matteo Pascucci (2021). “Deontic concepts and their clash in Mīmāṃsā: towards an interpretation”. In: *Theoria* 87.3, pp. 659–703. DOI: 10.1111/theo.12307.

Gabbay, Dov, Massimiliano Giacomin, Guillermo R. Simari, and Matthias Thimm (2021). *Handbook of Formal Argumentation, Volume 2*. United Kingdom: College Publications.

Gabbay, Dov, Ian Hodkinson, and Mark Reynolds (1994). *Temporal logic: mathematical foundations and computational aspects*. Oxford University Press, Oxford.

Gabbay, Dov, John F. Horty, Xavier Parent, Ron van der Meyden, and Leendert van der Torre (2013). *Handbook of Deontic Logic and Normative Systems, Volume 1*. United Kingdom: College Publications.

Gelfond, Michael, Vladimir Lifschitz, Halina Przymusinska, and Miroslaw Truszczynski (1991). "Disjunctive defaults". In: *Proceedings of the Second International Conference on Principles of Knowledge Representation and Reasoning*. Ed. by James Allen, Richard Fikes, and Erik Sandewall. Citeseer, pp. 230–237.

Gentzen, Gerhard (1934). "Untersuchungen über das logische Schließen I, II". In: *Mathematische Zeitschrift* 39, pp. 176–210, 405–431.

Giordani, Alessandro and Ilaria Canavotto (2016). "Basic Action Deontic Logic". In: *Deontic Logic and Normative Systems, 13th International Conference, (DEON 2016)*. Ed. by Olivier Roy, Allard M. Tamminga, and Malte Willer. College Publications, pp. 80–92.

Giordani, Alessandro and Matteo Pascucci (2022). "Generalizing Deontic Action Logic". In: *Studia Logica*, pp. 1–45.

Giunchiglia, Enrico, Joohyung Lee, Vladimir Lifschitz, Norman McCain, and Hudson Turner (2004). "Nonmonotonic causal theories". In: *Artificial Intelligence* 153.1-2, pp. 49–104.

Glavaničová, Daniela and Matteo Pascucci (2021). "The Good, the Bad and the Right: Formal Reductions among Deontic Concepts". In: *Bulletin of the Section of Logic* 50.2, pp. 151–176.

Goldman, Alvin I. (1970). *Theory of human action*. Princeton University Press.

Governatori, Guido and Mustafa Hashmi (2015). "No time for compliance". In: *2015 IEEE 19th International Enterprise Distributed Object Computing Conference*. IEEE, pp. 9–18.

Governatori, Guido and Antonino Rotolo (2006). "Logic of violations: A Gentzen system for reasoning with contrary-to-duty obligations". In: *The Australasian Journal of Logic* 4, pp. 193–215.

Governatori, Guido, Antonino Rotolo, and Régis Riveret (2018). "A deontic argumentation framework based on deontic defeasible logic". In: *International Conference on Principles and Practice of Multi-Agent Systems (PRIMA 2018)*. Ed. by Tim Miller, Nir Oren, Yuko Sakurai, Itsuki Noda, Bastin Tony Roy Savarimuthu, and Tran Cao Son. Springer, pp. 484–492.

Graham, Peter A. (2011). “’Ought’and Ability”. In: *Philosophical Review* 120.3, pp. 337–382.

Guhe, Eberhard (2021). “Killing Gently by Means of the Śyena: The Navya-Nyāya Analysis of Vedic and Secular Injunctions (vidhi) and Prohibitions (niṣedha) from the Perspective of Dynamic Deontic Logic”. In: *Journal of Indian Philosophy* 49 (3), pp. 421–449. DOI: `10.1007/s10781-021-09465-2`.

Hansson, Sven Ove (2013). “The varieties of permission”. In: *Handbook of deontic logic and normative systems, Volume 1* 1. Ed. by Dov Gabbay, John F. Horty, Xavier Parent, Ron van der Meyden, and Leendert van der Torre, pp. 195–240.

Hare, Richard M. (1971). “Practical inferences”. In: *Practical inferences*. University of California Press.

Herzig, Andreas and François Schwarzentruber (2008). “Properties of logics of individual and group agency.” In: *Advances in modal logic* 7. Ed. by Carlos Areces and Robert Goldblatt, pp. 133–149.

Hilpinen, Risto (1997). “On action and agency”. In: *Logic, action and cognition*. Springer, pp. 3–27.

Hilpinen, Risto and Paul McNamara (2013). “Deontic Logic: A Historical Survey and Introduction”. In: *Handbook of Deontic Logic and Normative Systems, Volume 1*. Ed. by Dov Gabbay, John F. Horty, Xavier Parent, Ron van der Meyden, and Leendert van der Torre. College Publications, pp. 3–136.

Hintikka, Jaakko (1970). “Some main problems of deontic logic”. In: *Deontic logic: Introductory and systematic readings*. Springer, pp. 59–104.

Horty, John F. (1994). “Moral Dilemmas and Nonmonotonic Logic”. In: *Journal of Philosophical Logic* 23.1, pp. 35–65. DOI: `10.1007/BF01417957`.

– (1997). “Nonmonotonic foundations for deontic logic”. In: *Defeasible deontic logic*. Ed. by Donald Nute. Springer, pp. 17–44.

– (2001). *Agency and deontic logic*. Oxford University Press.

– (2012). *Reasons as defaults*. Oxford University Press.

Horty, John F. and Nuel Belnap (1995). “The deliberative stit: A study of action, omission, ability, and obligation”. In: *Journal of philosophical logic* 24.6, pp. 583–644.

Hughes, Jesse, Peter Kroes, and Sjoerd Zwart (2007). “A semantics for means-end relations”. In: *Synthese* 158.2, pp. 207–231. DOI: 10.1007/s11229-006-9036-x.

Indrzejczak, Andrzej (1997). “Generalised sequent calculus for propositional modal logics.” In: *Logica Trianguli* 1, pp. 15–31.

Jakobovits, Hadassa and Dirk Vermeir (1999). “Robust semantics for argumentation frameworks”. In: *Journal of Logic and Computation* 9.2, pp. 215–261. DOI: 10.1093/logcom/9.2.215.

Johnson, Ralph H. (2000). *Manifest rationality: A pragmatic theory of argument*. Routledge.

Kanger, Stig (1971). “New foundations for ethical theory”. In: *Deontic logic: Introductory and systematic readings*. Ed. by Risto Hilpinen. Springer, pp. 36–58. DOI: 10.1007/978-94-010-3146-2_2.

– (1972). “Law and logic”. In: *Theoria* 38.3, pp. 105–132.

Kant, Immanuel (1998). *Critique of Pure Reason (The Cambridge Edition of the Works of Immanuel Kant)*. Ed. by Paul Guyer and Allen W. Wood. Cambridge University Press.

– (1999). “Groundwork for the Metaphysics of Morals”. In: *Practical Philosophy (The Cambridge Edition of the Works of Immanuel Kant)*. Ed. by Mary J. Gregor and Allen Wood. Cambridge University Press, pp. 37–108.

Kelsen, Hans (1991). *General theory of norms*. Ed. by Michael Hartney. Clarendon Press, Oxford.

Kohl, Markus (2015). “Kant and ‘ought implies can’”. In: *The Philosophical Quarterly* 65.261, pp. 690–710.

Kolodny, Niko and John MacFarlane (2010). “Ifs and oughts”. In: *The Journal of philosophy* 107.3, pp. 115–143.

Koons, Robert (2022). “Defeasible Reasoning”. In: *The Stanford Encyclopedia of Philosophy*. Ed. by Edward N. Zalta. Summer 2022. Metaphysics Research Lab, Stanford University.

Kratzer, Angelika (1981). “The notional category of modality”. In: *Words, worlds, and contexts* 38, pp. 38–74.

Lellmann, Björn (2021). “From Input/Output Logics to Conditional Logics via Sequents – with Provers”. In: *Automated Reasoning with Analytic Tableaux and Related Methods*. Ed. by Anupam Das and Sara Negri. Springer International Publishing, pp. 147–164.

Lellmann, Björn, Francesca Gulisano, and Agata Ciabattoni (2021). “Mīmāṃsā Deontic Reasoning using Specificity: a Proof Theoretic Approach”. In: *Artificial Intelligence and Law* 29, pp. 351–394. DOI: 10.1007/s10506-020-09278-w.

Lemmon, Edward John (1962). “Moral dilemmas”. In: *The philosophical review* 71.2, pp. 139–158.

Liao, Beishui, Nir Oren, Leendert van der Torre, and Serena Villata (2018). “Prioritized norms in formal argumentation”. In: *Journal of Logic and Computation* 29.2, pp. 215–240.

Liao, Beishui, Marija Slavkovik, and Leendert van der Torre (2019). “Building Jiminy Cricket: An Architecture for Moral Agreements Among Stakeholders”. In: *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society, AIES 2019*. Ed. by Vincent Conitzer, Gillian K. Hadfield, and Shannon Vallor. ACM, pp. 147–153.

Liao, Beishui and Leendert van der Torre (2020). “Explanation semantics for abstract argumentation”. In: *Frontiers in Artificial Intelligence and Applications, Computational Models of Argument*. Ed. by Henry Prakken, Stefano Bistarelli, Francesco Santini, and Carlo Taticch. Vol. 326. IOS Press, pp. 271–282. DOI: 10.3233/FAIA200511.

Lokhorst, Gert-Jan C. (1999). “Ernst Mally’s Deontik (1926)”. In: *Notre Dame Journal of Formal Logic* 40.2, pp. 273–282.

Lomuscio, Alessio and Marek Sergot (2003). “Deontic interpreted systems”. In: *Studia Logica* 75, pp. 63–92.

Lorini, Emiliano (2013). “Temporal logic and its application to normative reasoning”. In: *Journal of Applied Non-Classical Logics* 23.4, pp. 372–399.

Lorini, Emiliano and Giovanni Sartor (2014). “A STIT logic analysis of social influence”. In: *7th Workshop on Logical Aspects of Multi-Agent Systems (LAMAS 2014)*. Ed. by Nils Bulling and Wiebe van der Hoek, pp. 885–892.

– (2015). “Influence and Responsibility: A Logical Analysis”. In: *Frontiers in Artificial Intelligence and Applications: Legal Knowledge and Information Systems (JURIX 2015)*. Ed. by Antonio Rotolo. Vol. 279, pp. 51–60.

Lyon, Tim (2021). “Refining labelled systems for modal and constructive logics with applications”. PhD thesis. TU Wien.

Lyon, Tim and Kees van Berkel (2019). “Automating Agential Reasoning: Proof-Calculi and Syntactic Decidability for STIT Logics”. In: *PRIMA 2019: Principles and Practice of Multi-Agent Systems - 22nd International Conference*. Ed. by Matteo Baldoni, Mehdi Dastani, Beishui Liao, Yuko Sakurai, and Rym Zalila-Wenkstern. Vol. 11873. Lecture Notes in Computer Science. Springer, pp. 202–218. DOI: 10.1007/978-3-030-33792-6_13.

Makinson, David (1999). “On a fundamental problem of deontic logic”. In: *Norms, logics and information systems: New studies in deontic logic and computer science* 49, pp. 29–53.

Makinson, David and Leendert van der Torre (2000). “Input/Output logics”. In: *Journal of Philosophical Logic*, pp. 383–408.

– (2001). “Constraints for Input/Output Logics”. In: *Journal of Philosophical Logic* 30.2, pp. 155–185.

– (2003). “Permission from an Input/Output Perspective”. In: *Journal of Philosophical Logic* 32.4, pp. 391–416.

Marcus, Ruth Barcan (1980). “Moral dilemmas and consistency”. In: *The Journal of Philosophy* 77.3, pp. 121–136.

Markovich, Réka (2020). “Understanding Hohfeld and formalizing legal rights: the Hohfeldian conceptions and their conditional consequences”. In: *Studia Logica* 108.1, pp. 129–158.

McBurney, Peter and Simon Parsons (2009). “Dialogue games for agent argumentation”. In: *Argumentation in artificial intelligence*. Springer, pp. 261–280.

McConnell, Terrance (1985). “Metaethical principles, meta-prescriptions, and moral theories”. In: *American Philosophical Quarterly* 22.4, pp. 299–309.

– (1989). “‘Ought’ Implies ‘Can’ and the Scope of Moral Requirements”. In: *Philosophia* 19.4, pp. 437–454.

McCrea, Lawrence (2008). *The Teleology of Poetics in Medieval Kashmir*. Cambridge (Mass.): Department of Sanskrit and Indian Studies, Harvard University. DOI: 10.1017/S1479591410000173.

– (2010). “Hindu jurisprudence and scriptural hermeneutics”. In: *Hinduism and Law: An Introduction*. Ed. by Timothy Lubin, Jr. Davis Donald R., and Jayanth K. Krishnan. Cambridge: Cambridge University Press, pp. 123–136. DOI: 10.1017/CBO9780511781674.012.

Mercier, Hugo and Dan Sperber (2011). “Why do humans reason? Arguments for an argumentative theory”. In: *Behavioral and brain sciences* 34.2, pp. 57–74.

Meyer, John-Jules Ch. (1988). “A different approach to deontic logic: deontic logic viewed as a variant of dynamic logic.” In: *Notre Dame J. Formal Log.* 29.1, pp. 109–136. DOI: 10.1305/ndjfl/1093637776.

Meyer, John-Jules Ch., Jan Broersen, and Andreas Herzig (2015). “BDI Logics”. In: *Handbook of Logics of Knowledge and Belief*. Ed. by Hans van Ditmarsch, Joseph Y. Halpern, Wiebe van der Hoek, and Barteld Kooi. College Publications, pp. 453–498.

Meyer, John-Jules Ch., F. Dignum, and Roelf Johannes Wieringa (1994). *The Paradoxes of Deontic Logic Revisited: A Computer Science Perspective (Or: Should computer scientists be bothered by the concerns of philosophers?)* Tech. rep. UU-CS-1994-38. Department of Computer Sciences, Utrecht University.

Miller, Tim (2019). “Explanation in artificial intelligence: Insights from the social sciences”. In: *Artificial Intelligence* 267, pp. 1–38. DOI: https://doi.org/10.1016/j.artint.2018.07.007.

Modgil, Sanjay (2017). “Towards a General Framework for Dialogues That Accommodate Reasoning About Preferences”. In: *Theory and Applications of Formal Argumentation*, pp. 175–191.

Modgil, Sanjay and Martin Caminada (2009). “Proof theories and algorithms for abstract argumentation frameworks”. In: *Argumentation in artificial intelligence*. Springer, pp. 105–129.

Modgil, Sanjay and Henry Prakken (2013). “A general account of argumentation with preferences”. In: *Artificial Intelligence* 195, pp. 361–397.

– (2014). “The ASPIC+ framework for structured argumentation: a tutorial”. In: *Argument & Computation* 5.1, pp. 31–62.

Murakami, Yuko (2005). “Utilitarian deontic logic”. In: *Advances in Modal Logic* 5. Ed. by Renate Schmidt, Ian Pratt-Hartmann, Mark Reynolds, and Heinrich Wansing, pp. 287–302.

Nair, Shyam and John F. Horty (2018). “The Logic of Reasons”. In: *Oxford Handbooks Online*. Ed. by Daniel Star, pp. 67–84. DOI: 10.1093/oxfordhb/9780199657889.013.4.

Negri, Sara (2005). “Proof analysis in modal logic”. In: *Journal of Philosophical Logic* 34.5-6, pp. 507–544. DOI: 10.1007/s10992-005-2267-3.

– (2014). “Proof analysis beyond geometric theories: from rule systems to systems of rules”. In: *Journal of Logic and Computation* 26.2, pp. 513–537.

Negri, Sara and Edi Pavlović (2021). “Proof-theoretic analysis of the logics of agency: The deliberative STIT”. In: *Studia Logica* 109.3, pp. 473–507.

Negri, Sara, Jan Von Plato, and Aarne Ranta (2008). *Structural proof theory*. Cambridge University Press.

Nute, Donald (1997). *Defeasible deontic logic*. Vol. 263. Springer Science & Business Media.

Olkhovikov, Grigory K. and Heinrich Wansing (2018). “An axiomatic system and a tableau calculus for stit imagination logic”. In: *Journal of Philosophical Logic* 47.2, pp. 259–279.

Olszewski, Maya, Xavier Parent, and Leendert van der Torre (2021). “Input/Output Logic With a Consistency Check-the Case of Permission.” In: *Deontic Logic and Normative Systems - 15th International Conference, (DEON 2020/21)*. Ed. by Fenrong Liu, Alessandra Marra, Paul Portner, and Frederik Van De Putte. College publications, pp. 358–375.

Parent, Xavier (2011). “Moral particularism in the light of deontic logic”. In: *Artificial Intelligence and Law* 19(2), pp. 75–98.

– (2021). “Preference Semantics for Hansson-type Dyadic Deontic Logic: A Survey of Results”. In: *Handbook of deontic logic and normative systems, Volume 2*. Ed. by Dov Gabbay, John F. Horty, Xavier Parent, Ron van der Meyden, and Leendert van der Torre. College Publications, London, pp. 7–70.

Parent, Xavier and Leendert van der Torre (2013). “Input/Output Logic”. In: *Handbook of Deontic Logic and Normative Systems, Volume 1*. Ed. by Dov Gabbay, John F. Horty, Xavier Parent, Ron van der Meyden, and Leendert van der Torre. College Publications, pp. 499–544.

– (2017). “Detachment in normative systems: Examples, inference patterns, properties”. In: *IfCoLog Journal of Logics and Their Applications* 4.9, pp. 2295–3039.

– (2018a). *Introduction to deontic logic and normative systems*. College Publications.

– (2018b). “I/O Logics with a Consistency Check”. In: *Deontic Logic and Normative Systems, 14th International Conference, DEON 2018*. Ed. by Jan M. Broersen, Cleo Condoravdi, Nair Shyam, and Gabriella Pigozzi. College Publications, pp. 285–299.

Pascucci, Matteo (2017). “Anderson’s restriction of deontic modalities to contingent propositions”. In: *Theoria* 83.4, pp. 440–470.

Peirera, Célia da Costa, Andrea G.B. Tettamanzi, Serena Villata, Beishui Liao, Alessandra Malerba, Antonino Rotolo, and Leendert van der Torre (2017). “Handling norms in multi-agent system by means of formal argumentation”. In: *IfCoLog Journal of Logics and Their Applications* 4.9, pp. 1–35.

Pigozzi, Gabriella and Leendert van der Torre (2018). “Arguing about constitutive and regulative norms”. In: *Journal of Applied Non-Classical Logics* 28.2-3, pp. 189–217.

Pkhakadze, Sopo and Hans Tompits (2020). “Sequent-type calculi for three-valued and disjunctive default logic”. In: *Axioms* 9.3, pp. 1–29. DOI: 10.3390/axioms9030084.

Pollock, John L. (1987). “Defeasible reasoning”. In: *Cognitive science* 11.4, pp. 481–518.

Pörn, Ingmar (1977). *Action theory and social science*. Vol. 120. Springer Science & Business Media.

Prakken, Henry (2005). “Coherence and flexibility in dialogue games for argumentation”. In: *Journal of Logic and Computation* 15, pp. 1009–1040.

– (2018). “Historical overview of formal argumentation”. In: *Handbook of Formal Argumentation, Volume 1*. Ed. by Pietro Baroni, Dov Gabbay, Massimiliano Giacomin, Leendert van der Torre, et al. College Publications, pp. 75–143.

Prakken, Henry and Marek Sergot (1996). “Contrary-to-Duty Obligations”. In: *Studia Logica* 57.1, pp. 91–115. DOI: 10.1007/BF00370671.

Prior, Arthur (1967). *Past, present and future*. Oxford University Press.

Rao, Anand S. and Michael P. Georgeff (1995). “BDI agents: from theory to practice”. In: *ICMAS-95, Proceedings of the first international conference of multiagent systems*. Ed. by Viktor Lesser and Les Gasser. Vol. 95, pp. 312–319.

Rao, Anand S. and Michael P. Georgeff (1998). “Decision procedures for BDI logics”. In: *Journal of logic and computation* 8.3, pp. 293–343.

Reiter, Raymond (1980). “A logic for default reasoning”. In: *Artificial intelligence* 13.1-2, pp. 81–132.

Ross, Alf (1944). “Imperatives and Logic”. In: *Philosophy of Science* 11.1, pp. 30–46. DOI: doi:10.1086/286823.

Saka, Paul (2000). “Ought does not imply can”. In: *American Philosophical Quarterly* 37.2, pp. 93–105.

Saribatur, Zeynep G., Johannes P. Wallner, and Stefan Woltran (2020). “Explaining non-acceptability in abstract argumentation”. In: *Frontiers in Artificial Intelligence and Applications, 24th European Conference on Artificial Intelligence - ECAI 2020*. Ed. by Giuseppe De Giacomo, Alejandro Catala, Bistra Dilkina, Michela Milano, Senén Barro, Alberto Bugarín, and Jérôme Lang. Vol. 325. IOS Press, pp. 881–888.

Segerberg, Krister (1992). “Getting started: Beginnings in the logic of action”. In: *Studia logica*, pp. 347–378.

– (2002). “Outline of a logic of action”. In: *Advances In Modal Logic: Volume 3*. Ed. by Frank Wolter, Heinrich Wansing, Maarten de Rijke, and Michael Zakharyaschev. World Scientific, pp. 365–387.

Sergot, Marek (2008). “Action and Agency in Norm-Governed Multi-agent Systems”. In: *Engineering Societies in the Agents World VIII*. Ed. by Alexander Artikis, Gregory M. P. O’Hare, Kostas Stathis, and George Vouros. Springer Berlin Heidelberg, pp. 1–54.

Šešelja, Dunja and Christian Straßer (2013). “Abstract argumentation and explanation applied to scientific debates”. In: *Synthese* 190.12, pp. 2195–2217.

Shea-Blymyer, Colin and Houssam Abbas (2021). “Algorithmic Ethics: Formalization and Verification of Autonomous Vehicle Obligations”. In: *ACM Transactions on Cyber-Physical Systems (TCPS)* 5.4, pp. 1–25.

Stepin, Ilia, Jose M. Alonso, Alejandro Catala, and Martin Pereira-Farina (2021). “A survey of contrastive and counterfactual explanation generation methods for explainable artificial intelligence”. In: *IEEE Access* 9, pp. 11974–12001. DOI: 10.1109/ACCESS.2021.3051315.

Stocker, Michael (1971). “‘Ought’and ‘can”. In: *Australasian journal of philosophy* 49.3, pp. 303–316.

Stoutland, Frederick (2010). “Von Wright”. In: *A Companion to the Philosophy of Action*. Ed. by Timothy O’Connor and Constantine Sandis. John Wiley & Sons, pp. 589–597.

Straßer, Christian (2014). *Adaptive Logics for Defeasible Reasoning. Applications in Argumentation, Normative Reasoning and Default Reasoning*. Vol. 38. Trends in Logic. Springer.

Straßer, Christian and Aldo Antonelli (2019). “Non-monotonic Logic”. In: *The Stanford Encyclopedia of Philosophy*. Ed. by Edward N. Zalta. Summer 2019. Metaphysics Research Lab, Stanford University.

Straßer, Christian and Ofer Arieli (2014). “Sequent-Based Argumentation for Normative Reasoning”. In: *Deontic Logic and Normative Systems*. Ed. by Fabrizio Cariani, Davide Grossi, Joke Meheus, and Xavier Parent. Lecture Notes in Computer Science. Springer, pp. 224–240.

Straßer, Christian, Mathieu Beirlaen, and Frederik Van De Putte (2016). “Adaptive logic characterizations of input/output logic”. In: *Studia Logica* 104.5, pp. 869–916.

Straßer, Christian and Pere Pardo (2021). “Prioritized Defaults and Formal Argumentation”. In: *Deontic Logic and Normative Systems - 15th International Conference, (DEON 2021/21)*. Ed. by Fenrong Liu, Alessandra Marra, Paul Portner, and Frederik Van De Putte. College Publications, pp. 427–446.

Straßer, Christian and Ofer Arieli (2015). “Normative reasoning by sequent-based argumentation”. In: *Journal of Logic and Computation* 29.3, pp. 387–415. DOI: `10.1093/logcom/exv050`.

Subbāśāstrī, ed. (1929-1934). *Śrīmajjaiminipraṇitaṃ Mīmāṃsādarśanam*. Poona: Ānandāśramamudrāṇālaya.

Sun, Xin and Zohreh Baniasadi (2014). “STIT based deontic logics for the miners puzzle”. In: *12th European Conference on Multi-Agent Systems, EUMAS 2014*. Ed. by Nils Bulling. Springer, pp. 236–251.

Sverdlik, Steven (1985). “Counterexamples in ethics”. In: *Metaphilosophy* 16.2/3, pp. 130–145.

Thomason, Richmond H. (1981). “Deontic logic as founded on tense logic”. In: *New studies in deontic logic*. Ed. by Risto Hilpinen. Springer, pp. 165–176.

– (1984). “Combinations of tense and modality”. In: *Handbook of philosophical logic*. Ed. by Dov Gabbay and Franz Guenther. Springer, pp. 135–165.

Timmermann, Jens (2013). “Kantian dilemmas? Moral conflict in Kant’s ethical theory”. In: *Archiv für Geschichte der Philosophie* 95.1, pp. 36–64. DOI: 10.1515/agph-2013-0002.

van der Torre, Leendert (1997). “Reasoning about obligations: defeasibility in preference-based deontic logic”. PhD thesis. Erasmus University Rotterdam.

van der Torre, Leendert and Yao hua Tan (1998). “Deliberate Robbery, or the Calculating Samaritan”. In: *In Proceedings of the ECAI’98 Workshop on Practical Reasoning and Rationality (PRR’98)*.

Tosatto, Silvano Colombo, Guido Boella, Leendert van der Torre, and Serena Villata (2012). “Abstract normative systems: Semantics and proof theory”. In: *Proceedings of the 13th International Conference on Principles of Knowledge Representation and Reasoning (KR12)*. Ed. by Gerhard Brewka, Thomas Eiter, and Sheila A. McIlraith. AAAI Press, pp. 358–368.

Toulmin, Stephen E. (1958). *The Uses of Argument*. Cambridge University Press.

Trypuz, Robert and Piotr Kulicki (2015). “On deontic action logics based on Boolean algebra”. In: *Journal of Logic and Computation* 25.5, pp. 1241–1260.

Van Fraassen, Bas C. (1973). “Values and the heart’s command”. In: *The Journal of Philosophy* 70.1, pp. 5–19.

Vassiliades, Alexandros, Nick Bassiliades, and Theodore Patkos (2021). “Argumentation and explainable artificial intelligence: a survey”. In: *The Knowledge Engineering Review* 36, pp. 1–35. DOI: 10.1017/s0269888921000011.

Verhagen, Harko, Martin Neumann, and Munindar P. Singh (2018). “Normative multi-agent systems: Foundations and history”. In: *Handbook of normative multiagent systems*. Ed. by Amit Chopra, Leendert van der Torre, and Harko Verhagen. College Publications, pp. 3–25.

Viraraghavacharya, Uttamur T., ed. (1971). *Seśvaramīmāṃsā-Mīmāṃsāpaduke, Seswara Mīmāṃsā and Mīmāṃsā paduka [by Veṅkaṭanātha]*. Madras: Ubhaya Vedanta Granthamala.

Von Kutschera, Franz (1986). “Bewirken”. In: *Erkenntnis*, pp. 253–281.

– (1993). “Causation”. In: *Journal of philosophical logic* 22.6, pp. 563–588.

Vranas, Peter B.M. (2007). “I ought, therefore I can”. In: *Philosophical studies* 136.2, pp. 167–216.

– (2018a). “I ought, therefore I can obey”. In: *Philosopher’s Imprint* 18, pp. 1–36.

– (2018b). ““Ought” implies “can” but does not imply “must”: an asymmetry between becoming infeasible and becoming overridden”. In: *Philosophical Review* 127.4, pp. 487–514.

Walton, Douglas (2007). “Evaluating practical reasoning”. In: *Synthese* 157.2, pp. 197–240.

– (2010). “A dialogue system specification for explanation”. In: *Synthese* 182.3, pp. 349–374. DOI: 10.1007/s11229-010-9745-z.

Walton, Douglas and Chris Reed (2003). “Diagramming, argumentation schemes and critical questions”. In: *Anyone Who Has a View*. Springer, pp. 195–211.

Wansing, Heinrich (2006). “Tableaux for multi-agent deliberative-stit logic”. In: *Advances in modal logic* 6. Ed. by Guido Governatori, Ian Hodkinson, and Yde Venema, pp. 503–520.

von Wright, Georg Henrik (1951). “Deontic logic”. In: *Mind* 60.237, pp. 1–15. DOI: 10.1093/mind/LX.237.1.

– (1957). “The logical problem of induction”. In.

– (1963a). *Norm and action*. Routledge & Kegan Paul, fourth impression, London and Henley.

– (1963b). “Practical inference”. In: *The philosophical review* 72.2, pp. 159–179.

– (1968). *An essay in deontic logic and the general theory of action*. North-Holland Publishing Company, Amsterdam.

– (1972a). “On so-called practical inference”. In: *Acta sociologica* 15.1, pp. 39–53.

– (1972b). *The Varieties of Goodness*. Routledge & Kegan Paul, fourth impression, London and Henley.

– (1981). “On the logic of norms and actions”. In: *New studies in deontic logic*. Ed. by Risto Hilpinen. Springer, pp. 3–35.

Xu, Ming (1994a). “Decidability of deliberative stit theories with multiple agents”. In: *International Conference on Temporal Logic*. Springer, pp. 332–348.

Xu, Ming (1994b). “Decidability of stit theory with a single agent and Refref Equivalence”. In: *Studia Logica* 53.2, pp. 259–298.

– (1995). “On the basic logic of STIT with a single agent”. In: *The Journal of Symbolic Logic* 60.2, pp. 459–483.

– (1998). “Axioms for deliberative stit”. In: *Journal of Philosophical Logic* 27.5, pp. 505–552.

– (2010). “Combinations of stit and actions”. In: *Journal of Logic, Language and Information* 19, pp. 485–503.

– (2015). “Combinations of stit with ought and know”. In: *Journal of Philosophical Logic* 44.6, pp. 851–877.

Yaffe, Gideon (1999). “’Ought’ implies ’can’ and the principle of alternate possibilities”. In: *Analysis* 59.3, pp. 218–222.