Where Responsibility Takes You

Logics of Agency, Counterfactuals and Norms

Ilaria Canavotto

Where Responsibility Takes You

Logics of Agency, Counterfactuals and Norms

ILLC Dissertation Series DS-2020-16

Institute for Logic, Language and Computation

For further information about ILLC-publications, please contact

Institute for Logic, Language and Computation
Universiteit van Amsterdam
Science Park 107
1098 XG Amsterdam
phone: +31-20-525 6051
e-mail: illc@uva.nl
homepage: http://www.illc.uva.nl/

Cover design by Pietro Brugnetti.
Printed and bound by Ipskamp Printing.

ISBN: 978–94–6421–118–4

Where Responsibility Takes You

Logics of Agency, Counterfactuals and Norms

Academisch Proefschrift

ter verkrijging van de graad van doctor
aan de Universiteit van Amsterdam
op gezag van de Rector Magnificus
prof. dr. ir. K.I.J. Maex
ten overstaan van een door het College voor Promoties ingestelde
commissie, in het openbaar te verdedigen
op dinsdag 15 december 2020, te 12.00 uur

door

Ilaria Canavotto

geboren te Bergamo

Promotiecommisie

Promotores:	Prof. dr. F. Berto	Universiteit van Amsterdam
	Prof. dr. S.J.L. Smets	Universiteit van Amsterdam
Copromotor:	Prof. dr. A. Giordani	Università Cattolica del Sacro Cuore
Overige leden:	Prof. dr. J.F.A.K. van Benthem	Universiteit van Amsterdam
	Prof. dr. A. Betti	Universiteit van Amsterdam
	Dr. D. Grossi	Universiteit van Amsterdam
	Prof. dr. J.F. Horty	University of Maryland
	Prof. dr. F. Liu	Universiteit van Amsterdam
	Prof. dr. O. Roy	University of Bayreuth

Faculteit der Geesteswetenschappen

Ai miei genitori, Giovanni Canavotto e Rossella Zanini

Contents

Acknowledgments

First of all, I would like to thank my promotors, Francesco Berto and Sonja Smets, and my copromotor, Alessandro Giordani. I am grateful to Franz who enthusiastically supported my application to the UvA four years ago. Thank you, Franz, for insisting that I do both logic and philosophy, for introducing me to the logic and philosophy of imagination, and for helping me bring together the material on which this dissertation is based. I am grateful to Sonja for welcoming me so warmly into the Amsterdam Dynamics Group and for her invaluable guidance during this journey. Thank you, Sonja, for your enthusiasm and initiative, for always finding relevant connections between what I was doing and epistemic logic, and for teaching me how to make ambitious plans feasible. I am grateful to Alessandro who has supervised me on and off since I was an undergraduate student. Thank you, Alessandro, for the countless hours you spent answering my questions, for challenging my ideas, and for sharing your endless enthusiasm about research with me. I feel lucky to have had such a supportive and organized supervision team.

I would also like to thank Johan van Benthem, Arianna Betti, Davide Grossi, John Horty, Fenrong Liu, and Olivier Roy for accepting to be members of my doctoral committee. Your research has been a source of inspiration for me.

This dissertation would not have existed without the dedication and insights of the co-authors of the papers included in it: besides Franz, Sonja, and Alessandro, Alexandru Baltag and Eric Pacuit. I am grateful to Alexandru for contributing to each of our meetings with tons of ideas and for constantly pushing me to turn my abstract arguments into concrete examples. If the examples in Chapter 3 are about Alice killing Dan rather than Brutus killing Caesar (and the intricacies of Shakespeare's tragedy), it is entirely my fault. I am grateful to Eric for the many discussions we had about STIT, counterfactuals, and game theory, for his thought-provoking questions and inspiring comments, and for his contagious energy and excitement about all kinds of topics.

I would also like to thank Eric and John Horty for a wonderful visit at the

University of Maryland in fall, 2019. I am much indebted to Eric, Jeff, and the participants to the logic seminar – Casey Enos, Caleb Kendrick, Steven Kuhn, Paolo Santorio, Masayuki Tashiro, and Yichi Zhang – for the valuable feedback they gave me during my stay. I owe special thanks to Jeff for having been so supportive, both academically and personally, during and after my visit.

This thesis also benefited from insightful discussions and precious remarks I received at many conferences, seminars, and during some short visits.

Since the end of my Master's at the MCMP, I had the chance to attend the meetings of the PIOTR project, led by Olivier Roy and Piotr Kulicki. I learned a lot about deontic logic during these meetings. The comments I received from Olivier and Piotr, Albert Anglberger, Huimin Dong, Norbert Gratzl, Robert Trypuz, Davide Grossi, Dominik Klein, Alessandra Marra, Marek Sergot, and Frederik Van De Putte have very much contributed to my research.

Norbert Gratzl and Hannes Leitgeb gave me the opportunity to teach an intensive course around the topics of my dissertation at the MCMP in August, 2019. The second part of the thesis profited from the constructive inputs of the brave students who went through a whole week of deontic logic with Norbert and me. I would like to thank Norbert also for introducing me to deontic logic and, on a more personal note, for cheering me up with the best jokes.

I am grateful to Leon van der Torre and Réka Markovich for a very productive and inspiring two-day visit at the University of Luxembourg in March, 2019 and to Andrea Sereni and Maria Paola Sforza Fogliani for inviting me to present my work in Pavia in December, 2018.

The project that led to Chapter 3 originated from the vibrant discussions between the members of the Amsterdam Dynamics Group and the members of the Responsible Intelligent Systems Group led by Jan Broersen during the "STIT-DEL reading group" I organized together with Aldo Ramírez Abarca in 2017. I am grateful to the Utrecht group, especially Jan, Aldo, Hein Duijf, Alexandra Kuncová, and Allard Tamminga, for making me understand STIT and for their crucial feedback.

Doing a PhD at the ILLC gave me the opportunity to attend a lot of inspiring talks and regularly discuss my progress at weekly seminars. Two seminars, in particular, have been a constant along the way: At the Logic of Conceivability seminar, I learned all I know about imagination and epistemology of modality. I am truly grateful to Peter Hawke, Karolina Krzyżanowska, Aybüke Özgün, and Tom Schoonen both for all they taught me and for their extensive and accurate comments. The LIRa seminar contributed a great deal to my growth as a PhD student. I would like to thank Alexandru, Sonja, Nick Bezhanishvili, Malvin Gattinger, Aybüke Özgün, Soroush Rafiee Rad, Chenwei Shi, Anthia Solaki, Fernando Velázquez Quesada, Ana Lucía Vargas Sandoval, and Kaibo Xie for making organizing and participating in the seminar (and drinks!) such an enjoyable and stimulating experience. Finally, I would also like to acknowledge Tom Schoonen and Martin Lipman for having organized the Amsterdam Metaphysics Seminar

in 2017 and 2018, which kept my interest in metaphysics alive.

If I was able to organize so many travels and to deal with all practical matters, it was thanks to the help of Jenny Batson, Tanja Kassenaar, and Debbie Klaassen at the ILLC office and of Antoinette Allen and Louise Gilman at the Philosophy Department at the University of Maryland. I would also like to acknowledge the Amstardam University Fund for partly funding my research visit to the University of Maryland.

A warm word of thanks goes to my friends, who gave me strength in difficult moments and filled the last years with wonderful memories. In particular: Thank you, Anthi, Gianluca, Thom, and Tom, for being the best office-mates, for the many therapeutic beers, and for seeing to it that the mascara accident will never be forgotten. (And thank you, Tom, for translating my summary into Dutch!) Thank you, Aybüke and Seb, for taking care of me at home, for making every day a lot of fun, and for being such great flat-mates and friends. Thank you, Betta, Glo, Franci, and Michi, for having being there for me for so many years. Thank you, Pietro, for the many times you and Giulia hosted me in Milan and for designing the cover of this book.

Lastly, a very special "thank you" to my parents, Giovanni and Rossella, for believing in me more than anyone else, to my sister, Giulia, for teaching me to trust in myself, to my brother, Giorgio, and my sister-in-law, Giulia, for their constant encouragement, and to my uncle, Paolo, for his love and very precious cooking tips. I wouldn't have come this far without you!

Ilaria Canavotto

Amsterdam, October 2020

Chapter 1

Introduction

The purpose of this work is to bring together logics of agency, counterfactuals, and norms in order to address some key issues arising from a formal analysis of the notion of responsibility. We have in mind one of the most basic forms of responsibility, namely *causal responsibility*: responsibility deriving solely from the fact that a certain state of affairs has been brought about, no matter the intentions or beliefs of the agents involved. We design logical tools to start tackling three questions related to the development of a formal framework to reason about causal responsibility.

The first question is: *How can we model the agency of individuals and groups in causing certain results?* The question "Who is responsible for A?" often arises in situations involving a multiplicity of agents interacting in a multiplicity of ways. Only the agents that actually contributed to cause A are causally responsible for it. This calls for a formal framework with the resources to analyze *complex multiagent interactions* and, at the same time, capture elements of *causal reasoning*. In the last decades, computer scientists have developed powerful logical systems, such as Coalition Logic [Pauly, 2002] and Alternating-time Temporal Logic [Alur et al., 2002; Goranko and van Drimmelen, 2006], to reason about what interacting agents *can* do. However, these systems fall short in representing what the agents *actually* do, which is central to responsibility judgments. The most prominent logic of agency in the philosophical literature, namely STIT logic (i.e., the logic of *seeing to it that*) [Belnap et al., 2001; Horty, 2001], has the resources to model not only what interacting agents can do but also what they do *in fact*. Yet, STIT faces difficulties when applied to the analysis of cases in which the agents do not act independently of one another. The first contribution of the present work is to incorporate genuinely causal notions in STIT in order to overcome these difficulties. The improved framework will allow us to provide an analysis of causal responsibility based on considerations about potential as well as actual causality.

The second question we consider is: *What are the logical properties and epistemic value of counterfactuals concerning what can be, or could have been, done in*

the course of time? Causal responsibility is typically determined by considering what would have happened had the agents acted in a different way, where the agents often act sequentially. The formulation of responsibility judgments thus involves counterfactual statements about agents acting over time.

Regarding counterfactuals, a first problem we are concerned with is: *When are counterfactual statements true and when are they false*? The standard possible world semantics for counterfactuals due to Stalnaker [1968] and Lewis [1973a] abstracts away from considerations about time and agency. This led some scholars [Thomason and Gupta, 1981; Placek and Müller, 2007] to investigate the semantics of counterfactuals in the context of branching time – the theory of time that underlies STIT semantics. However, these proposals do not include a representation of agency, so counterfactual reasoning about what some of the agents would have done had others behaved differently cannot be fully captured or investigated in these frameworks. Our second contribution in this work is to address this issue by exploring the semantics and logical properties of counterfactuals in the context of STIT semantics.

But one thing is to define truth conditions for counterfactuals and another thing is to determine the epistemic value of the reasoning behind their evaluation. This takes us to a second problem concerning the kind of counterfactuals featuring in responsibility attributions: *How does the cognitive process that we use to evaluate them work? How does it generate knowledge?* There is a widespread agreement in philosophy and cognitive psychology that a distinctive mechanism underlying counterfactual reasoning is imagination, intended as reality oriented mental simulation [see, e.g., Williamson, 2007 in philosophy and Byrne and Girotto, 2009 in cognitive psychology]. The problematic point is that imagination seems to be voluntary in ways other mental states, like belief, are not: we can easily imagine that Alice has special bullets that pass through walls, while we can hardly make ourselves believe it, given overwhelming contrary evidence. But if, given some input, we can imagine anything we want, then imagination cannot lead to knowledge. The third contribution of this thesis is to advance a logical model of imagination in order to study its logic, its voluntary and involuntary components, and, relatedly, how it generates knowledge.

The last question we are interested in is: *Which rules govern normative reasoning?* Ascribing responsibility is not only a matter of identifying who caused a certain result, but also of determining which actions ought and ought not to be done: the question "Who is responsible?" only arises if something has been done that is *wrong* according to some moral or legal norms. A key issue here is how to reason about the wrongfulness of the actions (or sequences of actions) that can be performed in a given situation. Of course, an intuitive idea is that an action (or an action sequence) is wrong in a given situation if its performance leads to the violation of a norm. But what about situations in which a norm has already been violated? What if the circumstances make it impossible not to violate a norm again? Even worse, what if it is the normative system itself that requires

us to do things that cannot possibly be done together? What ought we do in such cases? And who is responsible for the resulting violations?

Although these questions (or variants thereof) have marked the history of deontic logic, they have mainly been investigated from a *static perspective*, i.e., by leaving aside considerations about whether, and how, the performance of an action (or a sequence of actions) can change the situation we are in. The last contribution of this dissertation is to investigate the previous issues from a *dynamic perspective*. We will do this by developing systems of deontic logic where actions, modeled as transitions from an initial-state (or model) to an end-state (or model), play a central role. Assuming this perspective will allow us to provide a fine-grained analysis of what it means that something is wrong in certain circumstances and to model different ways the agents may end up in a situation in which fulfilling all norms is impossible.

Overall, a characterizing feature of our contribution is the central role played by the notions of agency and action in the formal frameworks we advance. In this sense, the perspective from which we ask and answer the three aforementioned questions can be described as "*action-based*." As hinted above, we work within two modal traditions in the logic of action, namely STIT logic [Belnap et al., 2001; Horty, 2001] and dynamic logics, where the latter include Propositional Dynamic Logic [Harel et al., 2000] and Dynamic Epistemic Logic [Baltag et al., 1998; van Benthem, 2011; van Ditmarsch et al., 2008].

The thesis is organized as follows. Chapter 2 reviews the main logical systems that inspired our proposals. It introduces standard STIT semantics, overviews the main metalogical results, and compares STIT with Coalition Logic and Propositional Dynamic Logic. The chapter explains how these systems represent the notions of action and agency and what technical features characterize them.

The chapters in Part I (i.e., Chapters 3 to 5) address our first two questions, concerning the representation of the agency of individuals and groups in causing certain results and the logical and epistemic features of counterfactuals about agents acting in the course of time. Chapter 3 refines STIT logic by incorporating causal notions in it. We use the refined framework to formalize three key tests to ascribe causal responsibility, giving rise to three corresponding STIT operators, and to analyze ascriptions of individual and group responsibility in a number of examples. Chapter 4 extends the framework introduced in Chapter 3 and combines it with a logic of counterfactuals. We present three new STIT semantics for counterfactuals and discuss important philosophical and logical implications deriving from them. In Chapter 5, we use techniques from STIT logic, epistemic logic, and subject matter semantics to advance a logical model of imagination as reality oriented mental simulation. We address the question how such activity generates knowledge by investigating its logic and studying its voluntary and involuntary components.

The chapters in Part II (i.e., Chapters 6 and 7) address our last question, concerning normative reasoning. Chapter 6 presents a dynamic deontic logic

characterized by both a notion of ideality and a notion of optimality. We rely on it to provide a rich deontic classification of states, actions, and sequences of actions and to define deontic operators expressing so-called actual prescriptions – prescriptions that are sensitive to what can actually be done, given the circumstances. Actual prescriptions are of the greatest importance in situations in which the agents cannot avoid violating some norms. Chapter 7 zooms in on a main category of such situations, namely those resulting from the presence of a normative conflict. By borrowing ideas from explicit modal logics [Artemov, 2008; Fitting, 2005] and Dynamic Epistemic Logic, we design a framework to model the dynamics that gives rise to a conflict. We show how the resulting system can be used to keep track of the agents who generated a conflict and to capture distinctive aspects of cases of conscientious objection and civil disobedience.

Origin of the material

The chapters in Part I and Part II of this dissertation have either been published as articles, or are currently in preparation. Below I list the sources of the chapters and note the contribution of each author.

- Chapter 3 is based on the following article:

  > Alexandru Baltag, Ilaria Canavotto, and Sonja Smets. Causal agency and responsibility: A refinement of STIT logic. In Alessandro Giordani and Jacek Malinowski, editors, *Logic in High Definition, Trends in Logical Semantics*, volume 56 of *Trends in Logic*. Springer, Berlin. Forthcoming.

  Authors contributions: Alexandru Baltag and Sonja Smets initiated the project, Ilaria Canavotto developed the core motivation and applications and organized and coordinated the writing phase of the paper.

- Chapter 4 is based on the following paper:

  > Ilaria Canavotto and Eric Pacuit. Choice-driven counterfactuals. Manuscript in preparation. Institute for Logic, Language and Computation, University of Amsterdam and Department of Philosophy, University of Maryland.

  Authors contributions: the two authors discussed the central ideas and arguments together. Ilaria Canavotto organized and coordinated the writing phase of the paper.

- Chapter 5 is based on the following article:

  Ilaria Canavotto, Francesco Berto and Alessandro Giordani. Voluntary imagination: A fine-grained analysis. *The Review of Symbolic Logic*, pages 1-34, 2020.

  Authors contributions: Ilaria Canavotto and Francesco Berto initiated the project, Ilaria Canavotto and Alessandro Giordani developed the core ideas and Francesco Berto the philosophical motivation. The paper was co-written by the three authors.

- Chapter 6 extends the following article:

  Ilaria Canavotto and Alessandro Giordani. Erinching deontic logic. *Journal of Logic and Computation*, pages 241-263, 2019.

  which, in turn, develops ideas from the following paper:

  Alessandro Giordani and Ilaria Canavotto. Basic action deontic logic. In Olivier Roy, Allard Tamminga, and Willerd Malte, editors, *Deontic Logic and Normative Systems, 13th International Conference (DEON 2016)*, pages 80-92, College Publications, Milton Keynes. 2016.

  Authors contributions: the two authors contributed equally to the former paper, Alessandro Giordani initiated the latter paper.

- Chapter 7 is based on the following article:

  Ilaria Canavotto and Alessandro Giordani. Normative conflicts in a dynamic logic of norms and codes. In Jan M. Broersen, Cleo Condoravdi, Shyam Nair, and Gabriella Pigozzi, editors, *Deontic Logic and Normative Systems, 14th International Conference (DEON 2018)*, pages 71-90. College Publications, Milton Keynes. 2018.

  Authors contributions: the two authors contributed equally to the paper.

Chapter 2
Background on STIT and related logics

In this dissertation, we develop formal frameworks to reason about causal responsibility in multiagent scenarios and analyze how agency influences both counterfactual and normative reasoning. Our point of departure is the logic of *seeing to it that*, known by the acronym STIT [Belnap et al., 2001; Horty, 2001]. In the first part of this chapter [Sections 2.1 and 2.2], we provide a concise introduction to standard STIT semantics and overview the main metalogical results. In the second part [Section 2.3], we discuss the relation between STIT and two related logics for multiagent systems that have inspired the developments presented in subsequent chapters, namely Coalition Logic [Pauly, 2001, 2002] and Propositional Dynamic Logic [Harel et al., 2000]. The chapter aims at providing the background on how the basic notions of action and agency are represented in the aforementioned frameworks, and what the main conceptual and technical features of these systems are. In addition, the chapter also serves the purpose of fixing our notation and terminology for the two parts of the dissertation. The reader who is familiar with the topics should feel free to skim quickly through the definitions, or come back to them at their convenience later on. The reader who is not familiar with the topics but aims at a quick overview of the key concepts can focus on Section 2.1 and Section 2.3.2.

Preliminary remark. Although the chapter is introductory, we presuppose familiarity with basic modal logic. In particular, for all the logical systems that we consider here and in later chapters, the notions of theoremhood, deducibility, and consistency are defined in the standard way. The same applies to the notions of validity, logical consequence, and satisfiability. Finally, throughout the thesis, we will use the standard naming conventions for basic normal modal logics, like, for instance, K, KD, $\mathsf{S4}$, and $\mathsf{S5}$ [see Blackburn et al., 2001, Chapters 1 and 4].

2.1 Standard STIT semantics

STIT is a formal framework to reason about the agency of an individual or a group in bringing it about that, or *seeing to it that*, some state of affairs holds. The original theory, which stems from a modal tradition in the logic of action going back to St. Anselm and restarted in the 1960's by Alan Anderson, Brian Chellas, Fredric Fitch, Stig Kanger, and Franz von Kutschera among others,[1] was developed in a series of papers by Belnap, Perloff, and Xu, starting from Belnap and Perloff [1988] and culminating in Belnap et al. [2001]. The now standard extensions of STIT to groups and strategies are due to Horty [2001], who connected the original theory to issues in deontic logic and decision theory. Since Belnap et al. [2001] and Horty [2001], a number of different formulations, extensions, and applications of STIT have been studied, many of which will be mentioned later on in this and subsequent chapters. In this section, we present the standard theory. We start in Section 2.1.1 by gradually introducing the ingredients that are needed to give the semantics for the logic. Syntax and semantics will then be presented in Section 2.1.2.

2.1.1 Agency in branching time: STIT frames

Consider the following examples:

(1) There are fifty-two cards laying face up on the table. Alice picks the ace of spades, but she could have picked any of the other fifty-one cards.

(2) Max went for a run when the online seminar ended, but he could have joined an online yoga class instead.

(3) Giulia is driving from Bergamo to Milan. She takes the first exit towards Milan, but she could have waited and taken the second or the third exit.

Each of these examples can be represented as a STIT frame. STIT frames are based on the theory of branching time [Prior, 1967; Thomason, 1970, 1984] according to which the future can unfold in different ways, and how it will actually unfold depends, in part, on what the agents decide to do. *Branching time structures* (called *BT structures*) encode this view. A BT structure is a set of moments with a relation $<$ on this set, where $m < m'$ means that moment m occurs before moment m'. The relation $<$ is assumed to have a treelike structure (look at Figure 2.1) with forward branching representing the indeterminacy of the future and backward linearity representing the determinacy of the past.

2.1.1. Definition (BT structure). A BT structure is a tuple $\langle Mom, < \rangle$, where $Mom \neq \varnothing$ is a set of moments and $< \subseteq Mom \times Mom$ is the temporal precedence

[1] For concise historical overviews with the main references see Belnap et al. [2001, ch. 1D] and Segerberg [1992].

The picture depicts ten moments, m_1 to m_{10}, and six histories, h_1 to h_6. It is assumed that time flows upwards (hence, m_1 occurs, e.g., before m_2). Since all histories pass through m_1, $H_{m_1} = Hist$. So, for any history h_i, m_1/h_i is an index. Since only h_1 and h_2 pass through m_2, $H_{m_2} = \{h_1, h_2\}$. So, m_2/h_1 and m_2/h_2 are indices, while, e.g., m_2/h_3 is not.

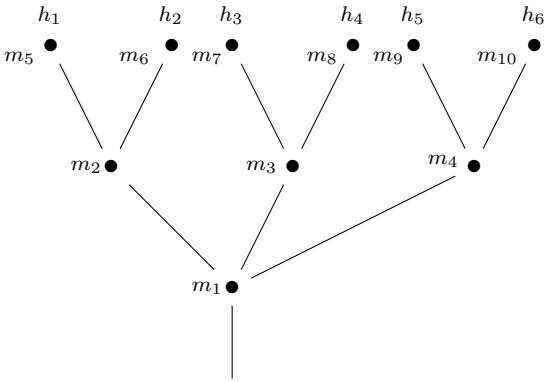


Figure 2.1: A branching time structure

relation between moments. As usual, $\leq \subseteq Mom \times Mom$ is defined as: for any $m, m' \in Mom$, $m \leq m'$ if and only if $m < m'$ or $m = m'$. The relation $<$ is assumed to satisfy the following conditions: for all $m, m_1, m_2, m_3 \in Mom$,

1. *Irreflexivity*: $m \not< m$.
2. *Transitivity*: if $m_1 < m_2$ and $m_2 < m_3$, then $m_1 < m_3$.
3. *Past-linearity*: if $m_1 \leq m_3$ and $m_2 \leq m_3$, then either $m_1 \leq m_2$ or $m_2 \leq m_1$.

A number of key notions, which are illustrated in Figure 2.1, can be defined in a BT structure $\mathcal{T}$. A *history* in $\mathcal{T}$ is a maximal set of linearly ordered moments from Mom, and represents a complete evolution of the world.[2] Let $Hist^{\mathcal{T}}$ be the set of all histories in $\mathcal{T}$. Because of forward branching, many different histories can pass through a single moment m (i.e., m can be an element of many different histories). The *set of histories passing through moment* m is $H_m^{\mathcal{T}} = \{h \in Hist^{\mathcal{T}} \mid m \in h\}$. The histories in H_m represent the possibilities that can still be realized at m, while the histories that do not pass through m can no longer be realized at m.[3] For instance, in example (1), when Alice picks a card, any of the fifty-two histories on which she picks a card on the table can be realized, but histories on which she is not playing cards cannot be realized. Similarly, in example (2), when the online seminar ends, the history on which Max goes for a run and the history on which he joins the online yoga class can both be realized; but histories on which Max did not attend the online seminar

[2] That is, a history h is a set of moments linearly ordered by $<$ such that if $h \subset h'$, then h' is not linearly ordered by $<$.

[3] When we say that a history represents a possibility we mean that it represents a possible complete course of events, not just a possible future. A possibility is open, or accessible, at m when the overall course of events it represents can still be realized at m.

The picture depicts two moments (m_1 and m_2) and five histories (h_1 to h_5). At m_1, the action available to agent 1 is $K_1 = \{h_1, h_2, h_3, h_4, h_5\}$ and the actions available to agent 2 are $K_2 = \{h_1\}$ and $K_3 = \{h_2, h_3, h_4, h_5\}$. At m_2, the actions available to agent 1 are $K_4 = \{h_3, h_4\}$ and $K_5 = \{h_2, h_5\}$, and the actions available to agent 2 are $K_6 = \{h_2, h_3\}$ and $K_7 = \{h_4, h_5\}$. The letters p and q display the valuation function defined in Example 2.1.11 below.

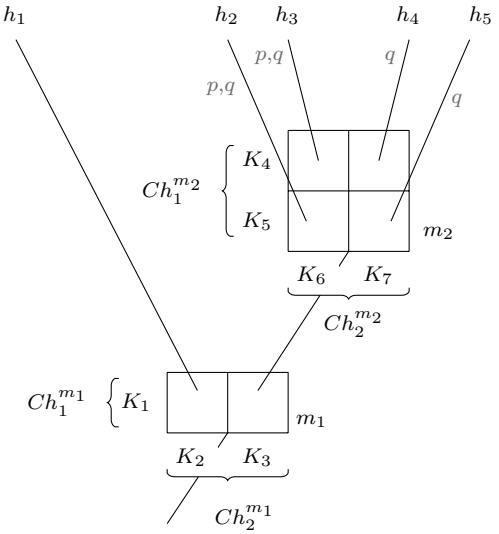


Figure 2.2: A STIT frame

cannot be realized. Example (3) can be understood in an analogous way. Finally, an *index* is any pair m/h such that $m \in Mom$ and $h \in H_m^{\mathcal{T}}$ and it intuitively corresponds to a complete state of the world at moment m on history h. As we will see later on, indices are the points of evaluation of STIT formulas. Let $Ind^{\mathcal{T}}$ be the set of all indices in $\mathcal{T}$. In what follows, we will omit the superscript $\mathcal{T}$ and simply write $Hist$, H_m and Ind when the BT structure is clear from the context.

STIT frames extend BT structures with descriptions of what the agents do at each moment. The main idea is that to act is to force the course of events to satisfy certain properties rather than others. When, e.g., Alice picks the ace of spades in example (1), she forces the history to be realized to be one on which she is holding the ace of spades, and she forces histories on which she is holding a different card to no longer be realizable.

This leads to a view where *the actions available to an agent i at a moment m* are represented (in a purely extensional way) as a partition of the histories passing through m: letting $Ag = \{1, \ldots, n\}$ be a fixed set of n (names of) agents for some number $n \in \mathbb{N}$, BT structures are supplemented with a function Ch that assigns to every agent i and moment m a partition Ch_i^m of the set of histories passing through m.[4] For example, in the STIT frame pictured in Figure 2.2, the histories passing through m_1 are $\{h_1, h_2, h_3, h_4, h_5\}$ and the agents' action sets are

[4]In STIT, agents are typically introduced as elements of STIT frames. Here we introduce them as elements of the syntax for technical convenience. We assume that the set Ag is finite because real-life situations never involve an infinite number of agents. This assumption will also simplify the formulation of some central axioms in the next section.

partitions of this set of histories: agent 1's action set at m_1 is the trivial partition $Ch_1^{m_1} = \{K_1\}$ where $K_1 = \{h_1, h_2, h_3, h_4, h_5\}$ and agent 2's action set at m_1 is the partition $Ch_2^{m_1} = \{K_2, K_3\}$ where $K_2 = \{h_1\}$ and $K_3 = \{h_2, h_3, h_4, h_5\}$. Given any index m/h, *the action that agent i performs at m/h* is the partition cell from agent i's action set containing h. We will also call this a *choice-cell* and denote it with $Ch_i^m(h)$.[5] For instance, in Figure 2.2, $Ch_2^{m_1}(h_3) = K_3 = \{h_2, h_3, h_4, h_5\}$ is the action that agent 2 performs at m_1/h_3. The histories in it are its *possible outcomes*. (Normally, a single action does not completely determine the future: for instance, Alice picking the ace of spades does not settle what card her opponent will pick or whether her neighbor will ring the doorbell.)

Note that every agent performs exactly one action at every index. Importantly, this action is an *action token*: a particular action occurring at a particular moment on a history. It is not an abstract, repeatable type of action. So, in a STIT frame representing, e.g., example (1), the set of histories associated with Alice's action represents Alice picking the particular ace of spade that is on the table at a specific moment, and not the general type of action *pick an ace of spade* [see Horty and Pacuit, 2017, p. 617]. We return to this issue below.

Besides representing actions as sets of histories, STIT is characterized by two assumptions about agency. The first, known as *no choice between undivided histories*, concerns the interaction between agency and branching time. According to it, no action available to any agent at a moment m can take apart histories that divide at some moment later than m. So, in the BT structure depicted in Figure 2.2, no action performed at m_1 can take apart h_2 and h_3: at m_1, it is not possible to exclude h_2 from the set of open possibilities without also excluding h_3, and vice versa. This means that the actions performed at m_1/h_2 and m_1/h_3 must be the same. The second assumption, known as *independence of agents*, concerns the interaction between different agents. According to it, any action available to any agent at a moment is compatible with any action available to any other agent at that moment. Hence, for any combination of actions available to different agents at a moment (one for each agent), there must be a history on which that combination of actions is performed at that moment. This means that, no matter what the agents separately decide to do, something will happen.[6] These notions are made precise by the following definitions.

2.1.2. Definition (Histories undivided at m). Let $\langle Mom, < \rangle$ be a BT structure and $m \in Mom$. Then, $h, h' \in Hist$ are undivided at m just in case $m \in h \cap h'$

[5] A terminological note is in order. In STIT, "action" and "choice" are used interchangeably to refer to the actions an agent can perform at a moment. Neither of these expressions refers to the intentions of the agent, or to the decision process leading to an action. Refinements of STIT accounting for these aspects of agency have been studied only recently: Herzig and Troquard [2006] is the first work that explores epistemic ideas in STIT, while Broersen [2011a] the first that introduces intentions.

[6] As we will see in more details in Section 2.2.2, this assumption implies that no agent can prevent another agent from performing any action available to her.

and there is $m' \in Mom$ such that $m < m'$ and $m' \in h \cap h'$.

2.1.3. Definition (BT choice structure). Let $Ag = \{1, \ldots, n\}$ be the set of (names of) agents defined above [see p. 10]. A BT choice structure is a tuple $\langle \mathcal{T}, Ch \rangle$ where $\mathcal{T}$ is a BT structure and $Ch : Ag \times Mom \to 2^{2^{Hist}}$ is a *choice function* that assigns to every $(i, m) \in Ag \times Mom$ a partition Ch_i^m of H_m.

2.1.4. Definition (Action selection function at m). Let $\langle \mathcal{T}, Ch \rangle$ be a BT choice structure and m a moment in $\mathcal{T}$. An action selection function at m is a mapping $s : Ag \to 2^{H_m}$ such that, for all $i \in Ag$, $s(i) \in Ch_i^m$. Sel_m is the set of all action selection functions at m.

Thus, an action selection function at m selects, for every agent i, an action available to i at m. For example, in Figure 2.2, the mapping $s : Ag \to 2^{H_{m_2}}$ such that $s(1) = K_5$ and $s(2) = K_6$ is an action selection function at m_2. So, an action selection function at m is basically a combination of actions available to the agents at m (one for each agent). Action selection functions are needed to state the condition of independence of agents.

2.1.5. Definition (STIT frame). A STIT frame is a BT choice structure $\langle \mathcal{T}, Ch \rangle$ satisfying the following conditions: for all $m \in Mom$, $h, h' \in Hist$, and $i \in Ag$,

1. *No Choice Between Undivided Histories*: if h and h' are undivided at m, then $h' \in Ch_i^m(h)$.
2. *Independence of Agents*: for all $s \in Sel_m$, $\bigcap_{i \in Ag} s(i) \neq \varnothing$.

As a final step, we extend the choice function Ch to groups of agents, where a group is any set of agents. The idea is that *an action available to a group at a moment* m is the intersection of some actions available to its members at m, one for each member. So, in Figure 2.2, the actions available to group $\{1, 2\}$ at m_2 are: $K_4 \cap K_6 = \{h_3\}$, $K_4 \cap K_7 = \{h_4\}$, $K_5 \cap K_6 = \{h_2\}$, and $K_5 \cap K_7 = \{h_5\}$.

2.1.6. Definition (Group choice). Let $\langle \mathcal{T}, Ch \rangle$ be a STIT frame. For any group of agents $I \subseteq Ag$ and moment $m \in Mom$, the set of actions available to I at m is defined as: $Ch_I^m = \{\bigcap_{i \in I} s(i) \mid s \in Sel_m\}$.

The condition of independence of agents ensures that group choices are never empty – in fact, it is easy to see that Ch_I^m is a partition of H_m. In addition, given Definition 2.1.6, the larger a group I is the finer the partition Ch_I^m will be, and so the greater I's control on the future will be. Accordingly, for every moment m, Ch_{Ag}^m is the finest partition on H_m, while $Ch_{\varnothing}^m = \{H_m\}$ the coarsest.

2.1.2 Syntax and Semantics

We now introduce the language $\mathcal{L}_{\mathsf{STIT}^G_n}$ of the logic STIT^G_n of *group temporal STIT* and two fragments of this language that will be important in the next section. We start by fixing, besides the set $Ag = \{1, \ldots, n\}$ of (names of) agents [see p. 10], a non-empty countable set $Prop$ of propositional variables. We will use $i, j, k, i', i'', \ldots$ for elements of Ag and $p, q, r, p', p'', \ldots$ for elements of $Prop$.

2.1.7. Definition (Syntax of $\mathcal{L}_{\mathsf{STIT}^G_n}$). Let Ag and $Prop$ be defined as above. The set of formulas of $\mathcal{L}_{\mathsf{STIT}^G_n}$, also denoted with $\mathcal{L}_{\mathsf{STIT}^G_n}$, is generated by the following grammar:

$$\varphi := p \mid \neg\varphi \mid (\varphi \wedge \varphi) \mid \mathsf{G}\varphi \mid \mathsf{H}\varphi \mid \Box\varphi \mid [I\, cstit]\varphi$$

where $p \in Prop$ and $I \subseteq Ag$. The abbreviations for the Boolean connectives $\vee$, $\rightarrow$, and $\leftrightarrow$, for $\top$ and $\bot$ are standard. As usual, we use $\mathsf{F}\varphi$, $\mathsf{P}\varphi$, $\Diamond\varphi$, and $\langle I\, cstit\rangle\, \varphi$ as abbreviations for $\neg\mathsf{G}\neg\varphi$, $\neg\mathsf{H}\neg\varphi$, $\neg\Box\neg\varphi$, $\neg[I\, cstit]\neg\varphi$. In addition, we write $[i\, cstit]\varphi$ instead of $[\{i\}\, cstit]\varphi$. Finally, we will adopt the usual rules for the elimination of the parentheses.

2.1.8. Definition (Fragments of $\mathcal{L}_{\mathsf{STIT}^G_n}$). The language $\mathcal{L}_{\mathsf{A\text{-}STIT}^G_n}$ of *group atemporal STIT*, is the fragment of $\mathcal{L}_{\mathsf{STIT}^G_n}$ without the temporal modalities G and H. The language $\mathcal{L}_{\mathsf{A\text{-}STIT}_n}$ of *individual atemporal STIT* is the fragment of $\mathcal{L}_{\mathsf{A\text{-}STIT}^G_n}$ without the group modalities $[I\, cstit]$, where $|I| > 1$.

The full language $\mathcal{L}_{\mathsf{STIT}^G_n}$ of group temporal STIT includes three types of modalities: the usual temporal operators G and H of, respectively, past necessity and future necessity on a history, the operator $\Box$ of historical necessity, and, finally, the so-called Chellas-STIT operators $[I\, cstit]$.[7] The intended interpretation of the modal formulas is as follows. $\mathsf{G}\varphi$ means "φ will always be true in the future" and $\mathsf{H}\varphi$ means "φ has always been true in the past." These formulas are taken to be true at an index m/h whenever φ is true, respectively, at all indices m'/h such that $m < m'$ and at all indices m''/h such that $m'' < m$. Next, $\Box\varphi$ means "φ is settled true" or "φ is historically necessary." A formula like $\Box\varphi$ is assumed to be true at an index m/h whenever φ turns out to be true at m no matter how the future unfolds from m on. Finally, $[I\, cstit]\varphi$ says "group I sees to it that φ." The semantics for this is based on the idea that an agent sees to it that φ just in case what that agent does ensures that φ obtains, no matter what the other agents do. Accordingly, $[I\, cstit]\varphi$ is taken to be true at an index m/h just in case φ is true at m on all histories that might result from the action performed by I at m/h.

The notions of STIT model and truth of formulas from $\mathcal{L}_{\mathsf{STIT}^G_n}$ at an index are defined as follows.

[7] The operator was named by Horty and Belnap [1995] after Brian Chellas, because it is an analogue of the operator introduced in Chellas [1969]. Chellas compares his work on agency and STIT in Chellas [1992], but see also Horty and Belnap [1995].

2.1.9. Definition (STIT model). Let *Prop* be defined as above. A STIT model is a tuple $\langle \mathcal{F}, \pi \rangle$ where $\mathcal{F}$ is a STIT frame and $\pi : Prop \to 2^{Ind}$ is a valuation function assigning to every propositional variable the set of indices where it is true.

2.1.10. Definition (STIT semantics for $\mathcal{L}_{\mathsf{STIT}_n^G}$). Given a STIT model $\mathcal{M}$, truth of a formula $\varphi \in \mathcal{L}_{\mathsf{STIT}_n^G}$ at an index m/h in $\mathcal{M}$, denoted $\mathcal{M}, m/h \models \varphi$, is defined recursively as follows:

$$
\begin{array}{lll}
\mathcal{M}, m/h \models p & \text{iff} & m/h \in \pi(p) \\
\mathcal{M}, m/h \models \neg\varphi & \text{iff} & \mathcal{M}, m/h \not\models \varphi \\
\mathcal{M}, m/h \models \varphi \wedge \psi & \text{iff} & \mathcal{M}, m/h \models \varphi \text{ and } \mathcal{M}, m/h \models \psi \\
\mathcal{M}, m/h \models \mathsf{G}\varphi & \text{iff} & \text{for all } m' \in h, \text{ if } m < m' \text{ then } \mathcal{M}, m'/h \models \varphi \\
\mathcal{M}, m/h \models \mathsf{H}\varphi & \text{iff} & \text{for all } m' \in h, \text{ if } m' < m \text{ then } \mathcal{M}, m'/h \models \varphi \\
\mathcal{M}, m/h \models \Box\varphi & \text{iff} & \text{for all } h' \in H_m,\ \mathcal{M}, m/h' \models \varphi \\
\mathcal{M}, m/h \models [I\, cstit]\varphi & \text{iff} & \text{for all } h' \in Ch_I^m(h),\ \mathcal{M}, m/h' \models \varphi
\end{array}
$$

2.1.11. Example. Consider index m_2/h_2 in the STIT frame pictured in Figure 2.2. Let $\pi(p) = \{m_2/h_2, m_2/h_3\}$ and $\pi(q) = \{m_2/h_2, m_2/h_3, m_2/h_4, m_2/h_5\}$. Since q is true at m_2 on all histories passing through it, $\Box q$ is true at m_2/h_2. Since p is true at m_2 on some but not all histories passing through it, $\Diamond p$ is true at m_2/h_2, whereas $\Box p$ is not. In addition, since all histories in $Ch_1^{m_2}(h_2) = K_6 = \{h_2, h_3\}$ are such that p and q are true at m_2 on these histories, both $[1\, cstit]p$ and $[1\, cstit]q$ are true at m_2/h_2. Similarly, since q is true at m_2 on all histories in $Ch_2^{m_2}(h_2) = K_5 = \{h_2, h_5\}$, $[2\, cstit]q$ is true at m_2/h_2. Yet, as p is false at m_2/h_5, $[2\, cstit]p$ is false at m_2/h_2.

By relying on Example 2.1.11, let us conclude this section with some remarks on the treatment of three important aspects of agency in STIT.

Temporal gap between an action and its effects. Example 2.1.11 shows that there is no time lapse between the fact that, say, agent 1 performs action K_6 and the fact that its effects, p and q, obtain: these facts all happen at the same moment m_2. Other operators have been discussed in the literature, notably the achievement STIT operator [Belnap and Perloff, 1988] and the XSTIT operator [Broersen, 2009, 2011b], which are based on the assumption that it takes time for an action to produce its effects. However, operators accounting for this assumption can also be introduced in $\mathcal{L}_{\mathsf{STIT}_n^G}$ by simply combining the Chellas-STIT operator $[I\, cstit]$ with the temporal operator G or its dual F: formulas like $[I\, cstit]\mathsf{G}\varphi$ and $[I\, cstit]\mathsf{F}\varphi$ say that φ is a future effect of what group I is doing. "Future oriented" versions of the Chellas-STIT operator can then be defined in $\mathcal{L}_{\mathsf{STIT}_n^G}$ by using these formulas. Importantly, this does *not* mean that STIT can represent actions as having a duration: by modeling them as sets of histories passing through the moment at which the agents decide to act, STIT semantics abstracts away from the dynamic dimension of actions.

Deliberativeness. Another interesting feature of Example 2.1.11 is that at m_2 the agents bring about q, which is settled true at that moment. But in what sense can an agent be said to bring about something whose realization is guaranteed, no matter what she will do? The so-called deliberative STIT operator, first proposed by von Kutschera [1986] and independently suggested by Horty [1989], encodes the intuitive judgment that there is no sense in which an agent can be said to bring about what cannot be otherwise.[8] This operator can be introduced by definition in $\mathcal{L}_{\mathsf{STIT}_n^G}$ as follows.

2.1.12. Definition (Deliberative STIT). Where $\varphi \in \mathcal{L}_{\mathsf{STIT}_n^G}$ and $I \subseteq Ag$,

$$[I\, dstit]\varphi := [I\, cstit]\varphi \wedge \neg\Box\varphi$$

The relation between Chellas and deliberative STIT operators is extensively investigated in Horty and Belnap [1995]. One important point is that, in the presence of the historical necessity operator, it is also possible to define the Chellas-STIT operator from the deliberative STIT operator by setting: $[I\, cstit]\varphi := [I\, dstit]\varphi \vee \Box\varphi$. Since the two operators are interdefinable, it is immaterial which one of the two is assumed as primitive. As it will become apparent in the next section, the Chellas-STIT operator (we will call it simply "STIT operator" from now on) is typically preferred because of its technical convenience.

Ability. Finally, it is worth noting that, in Example 2.1.11, there is an asymmetry between the actions available to the two agents at m_2. We have seen above that, unlike agent 1, agent 2 does not see to it that p at m_2/h_2. But the asymmetry is deeper than this: at m_2, there is nothing that agent 2 can do to make p true. According to a view that goes back to Horty and Belnap [1995], this means that agent 2 lacks the *ability* to see to it that p. In general, the assumption is that an agent i has the ability to see to it that φ when it is (historically) possible that i sees to it that φ, i.e., when $\Diamond[i\, cstit]\varphi$ holds. By applying Definition 2.1.10, it is not difficult to see that the latter formula has the following truth condition:

$$\begin{aligned}\mathcal{M}, m/h \models \Diamond[i\, cstit]\varphi \text{ iff } &\text{there is } K \in Ch_i^m \text{ such that,}\\ &\text{for all } h' \in K,\ \mathcal{M}, m/h' \models \varphi\end{aligned}$$

Thus, an agent has the ability to see to it that φ whenever there is an action available to her that guarantees the truth of φ. Of particular interest in the truth condition is the $\exists\forall$ pattern of the quantifiers, which makes explicit the twofold character of the notion of ability: ability involves not only potentiality (the "there

[8] As explained by Horty and Belnap [1995], the terminology goes back to Aristotle's remark that we can properly be said to deliberate only about "what is future and capable of being otherwise" [*Nichomachean Ethics*, 1139b7, and 1112a19-b10]. This idea is also common in modern semantics [see, e.g., Perry, 1989].

is" part) but also control on the result (the "for all" part).[9] We will come back to this when we compare STIT to other frameworks. Before that, it is now time to dive more deeply into STIT logic and its metalogical properties.

2.2 Metalogical results and Kripke semantics

While STIT has played a prominent role in the philosophical literature since the 1980's, a full exploration of STIT logic only started in recent years. Central in this process has been the connection of standard STIT semantics with better understood frameworks, especially Kripke semantics. In this section, we review the main metalogical results and the Kripke semantics for STIT. Following the history of the field, we start from *atemporal* STIT, i.e., the fragment of STIT without temporal operators.

2.2.1 Axiomatization of individual atemporal STIT

The logic of *individual* atemporal STIT was first axiomatized by Xu [1995, 1998], who later proved decidability using a filtration argument [see Belnap et al., 2001, Chapter 17]. More recently, Wölfl [2002] gave an alternative axiomatization by extending the language with extra modal operators, and Wansing [2006] provided a complete tableaux calculus. Two further axiomatizations, equivalent to the one proposed by Xu, were presented by Balbiani et al. [2008b]. We define the axiom system $\mathsf{A\text{-}STIT}_n$ by assuming Xu's, by now standard, axiomatization.

2.2.1. Definition ($\mathsf{A\text{-}STIT}_n$). The axiom system $\mathsf{A\text{-}STIT}_n$ of individual atemporal STIT is defined by the axioms and rules in Table 2.1.

(CPL)	All classical propositional tautologies
($\mathsf{S5}_\Box$)	The axiom schemas of S5 for $\Box$
($\mathsf{S5}_{[i\,cstit]}$)	The axiom schemas of S5 for $[i\,cstit]$
(Inc)	$\Box\varphi \to [i\,cstit]\varphi$
(IA)	$(\Diamond[1\,cstit]\varphi_1 \wedge \cdots \wedge \Diamond[n\,cstit]\varphi_n) \to \Diamond([1\,cstit]\varphi_1 \wedge \cdots \wedge [n\,cstit]\varphi_n)$
(MP)	From φ and $\varphi \to \psi$, infer ψ
($\mathsf{RN}_\Box$)	From φ, infer $\Box\varphi$

Table 2.1: The axiom system $\mathsf{A\text{-}STIT}_n$

[9] The $\exists\forall$ pattern also characterizes the logic of ability proposed by Brown [1988]. The connection between Brown's proposal and the analysis of ability in STIT is again explored in Horty and Belnap [1995], where the authors also address Anthony Kenny's [1975; 1976] well known arguments that ability is not a kind of possibility.

2.2.2. Theorem. [Belnap et al., 2001, Chapter 17]. *The axiom system* $\mathsf{A\text{-}STIT}_n$ *is sound and complete with respect to the class of all STIT frames.*

A short detour into Kripke semantics for atemporal STIT will make clear why $\mathsf{A\text{-}STIT}_n$ has the axioms it has. This alternative semantics, which is implicit in Xu's [2001] completeness proof, was introduced by Balbiani et al. [2008b], with the specific aim of exploring the mathematical properties of STIT, and, independently, by Kooi and Tamminga [2008], with the aim of generalizing Horty's [2001] utilitarian deontic logic to study moral conflicts in a multiagent setting. While Balbiani et al. [2008b] focus on *individual* atemporal STIT, Kooi and Tamminga [2008] base their analysis on *group* atemporal STIT. The main ideas of the semantics can be summarized as follows.

As usual, we start from a set W of possible states. States in W represent *moment-history pairs*, i.e., the points of evaluation in standard STIT semantics. We partition W by grouping together states that, intuitively, stand for indices consisting of the same moment. Each cell in this partition (call it the *moment-partition*) represents a *moment*, namely the one that grounds the grouping of the states in the cell. Given this interpretation, different states in the same cell stand for indices built from different *histories passing through a moment* (the one represented by the cell). Every state in the cell thus witnesses, so to speak, a history passing through the corresponding moment. This makes it natural to represent the set of *actions available to an agent at a moment* as a partition of the cell representing that moment. If we zoom out and look at all moments at once, the result is that every agent is associated with a partition of the set of all possible states (call it the *agent-partition*) that refines the moment-partition (see Figure 2.3 on page 22 below for an illustration).

In the framework of *atemporal* STIT the focus is restricted to agents acting at a single moment of time, while the full temporal evolution leading to, and following, this moment is abstracted away. So, the moment-partition is the trivial partition having W as the only element – and can thus be identified with this set. Each agent-partition represents the actions available to an agent at the single moment represented by W. This leads to the following notion of Kripke STIT frame. Notice that, in line with standard Kripke semantics for modal logic, in the next definition we use equivalence relations instead of partitions. In addition, for any binary relation R on a set X and $x \in X$, we define $R(x) = \{x' \in X \mid xRx'\}$.

2.2.3. Definition (Kripke A-STIT frame). A Kripke A-STIT frame is a tuple $\langle W, R\rangle$ where $W \neq \varnothing$ is a set of possible states and $R : Ag \to 2^{W\times W}$ assigns to every agent i an equivalence relation R_i on W. For any $w \in W$, $R_i(w)$ is the action performed by i at w. The map R is assumed to satisfy:

1. *Independence of Agents*: for all $w_1, \ldots, w_n \in W$, $\bigcap_{i\in Ag} R_i(w_i) \neq \varnothing$.

As before, the condition of independence of agents expresses that any action available to any agent must be compatible with any action available to any

other agent. As in standard STIT semantics, group actions are intersections of individual actions:

2.2.4. Definition (Group choices). Let $\langle W, R\rangle$ be a Kripke A-STIT frame. For any $I \subseteq Ag$, we set: $R_I = \bigcap_{i\in I} R_i$.

2.2.5. Definition (Kripke A-STIT model). Let *Prop* be defined as above. A Kripke A-STIT model is a tuple $\langle F, \nu\rangle$ where F is a Kripke A-STIT frame and $\nu : Prop \to \wp(W)$ is a valuation function.

Conceptually, the next definition is a natural consequence of viewing each state in a Kripke A-STIT model $\langle W, R, \nu\rangle$ as a (witness for a) history passing through the single moment represented by W.

2.2.6. Definition (Kripke semantics for $\mathcal{L}_{\mathsf{A\text{-}STIT}^G_n}$). Given a Kripke A-STIT model M, truth of a formula $\varphi \in \mathcal{L}_{\mathsf{A\text{-}STIT}^G_n}$ at a state w in M, denoted $M, w \models \varphi$, is defined recursively as follows:

$$
\begin{array}{lll}
M, w \models p & \text{iff} & w \in \nu(p) \\
M, w \models \neg\varphi & \text{iff} & M, w \not\models \varphi \\
M, w \models \varphi \wedge \psi & \text{iff} & M, w \models \varphi \text{ and } M, w \models \psi \\
M, w \models \Box\varphi & \text{iff} & \text{for all } w' \in W,\ M, w \models \varphi \\
M, w \models [I\, cstit]\varphi & \text{iff} & \text{for all } w' \in W, \text{ if } wR_Iw', \text{ then } M, w' \models \varphi
\end{array}
$$

The next theorem can now be proved by applying standard techniques from modal logic [see Blackburn et al., 2001, Chapter 4.2].

2.2.7. Theorem. *The axiom system* $\mathsf{A\text{-}STIT}_n$ *is sound and complete with respect to the class of all Kripke A-STIT frames.*

The precise connection between the standard and the Kripke semantics for atemporal STIT is established by the following result, whose proof is based on the above-mentioned interpretation of Kripke A-STIT frames.

2.2.8. Theorem. [Herzig and Schwarzentruber, 2008] *For every formula* $\varphi \in \mathcal{L}_{\mathsf{A\text{-}STIT}^G_n}$, φ *is satisfiable in a Kripke A-STIT model just in case it is satisfiable in a STIT model.*

Completeness of $\mathsf{A\text{-}STIT}_n$ with respect to the class of all STIT frames now follows as an immediate corollary of Theorems 2.2.7 and 2.2.8. In addition, our detour gives us a new understanding of standard STIT semantics, which makes the intuitiveness of the axioms of $\mathsf{A\text{-}STIT}_n$ apparent: The axioms of the modal logic $\mathsf{S5}$ for $\Box$ and $[i\, cstit]$ express that historical necessity and necessity resulting from acting are modeled as equivalence relations. Axiom Inc reflects the fact that each agent-partition refines the moment-partition, or, equivalently, that the actions available to an agent at a moment are modeled as subsets of the set of

histories passing through that moment. Finally IA, the axiom for independence of agents, expresses that, at any moment, the intersection of any combination of actions available to the agents, one for each agent, must be non-empty. Clearly, the absence of an axiom for the condition of no choice between undivided histories depends on the fact that the temporal dimension has been abstracted away.

2.2.2 Independence of agents and complexity

Axiom IA is a central axiom of STIT logic, both from a conceptual and from a technical point of view. As proved by Balbiani et al. [2008b], this axiom can be replaced either with schema IA' or with the union of schemas IA'' and IA''' below:

(IA') $\quad \Diamond\varphi \rightarrow \langle i\, cstit\rangle \bigwedge_{j\in Ag\setminus\{i\}} \langle j\, cstit\rangle\, \varphi$

(IA'') $\quad \Box\varphi \leftrightarrow [i\, cstit][j\, cstit]\varphi$

(IA''') $\quad \Diamond\varphi \rightarrow \langle k\, cstit\rangle \bigwedge_{i\in I\setminus\{k\}} \langle i\, cstit\rangle\, \varphi$ where $I \subseteq Ag$

From a conceptual point of view, IA', IA'', and IA''' describe the power that an agent can exercise over other agents. Unsurprisingly, these principles limit such power: According to IA' and IA''', no agent can see to it that another agent prevents a possible state of affairs from happening. According to IA'', the only states of affairs an agent can guarantee that another agent brings about are those that are already settled. This reveals that independence of agents is a strong assumption, which excludes the possibility of representing, in a non-trivial way, agents making other agents do something or prevent something from happening.

From a technical point of view, as observed by Balbiani et al. [2008b], in case $|Ag| = 2$, IA'' and IA''' ensure the derivability of the permutation axiom

$$\langle i\, cstit\rangle\, \langle j\, cstit\rangle\, \varphi \leftrightarrow \langle j\, cstit\rangle\, \langle i\, cstit\rangle\, \varphi$$

as well as of the Church-Rosser axiom

$$\langle i\, cstit\rangle\, [j\, cstit]\varphi \rightarrow [j\, cstit]\, \langle i\, cstit\rangle\, \varphi$$

This tells us that, in case Ag has only two agents, the logic $\mathsf{A\text{-}STIT}_n$ is nothing but the product logic $\mathsf{S5} \otimes \mathsf{S5}$. By applying results about the latter logic [see Marx, 1999], we conclude that the satisfiability problem for $\mathcal{L}_{\mathsf{A\text{-}STIT}_n}$ with two agents is NEXPTIME complete. As shown by Balbiani et al. [2008b], this remains true for any number of agents greater than 2. On the other hand, for the single-agent case, $\mathsf{A\text{-}STIT}_n$ has the same complexity as $\mathsf{S5}$.

2.2.9. Theorem. [Balbiani et al., 2008b] *The problem of deciding satisfiability of a formula of $\mathcal{L}_{\mathsf{A\text{-}STIT}_n}$ is: NP complete if $n = 1$; NEXPTIME complete if $n \geq 2$.*

So, all in all, $\mathsf{A\text{-}STIT}_n$ has very convenient formal properties. Unfortunately, however, these results do not extend to group (atemporal) STIT:

($\mathsf{A\text{-}STIT}_n$)	Axioms and rules of $\mathsf{A\text{-}STIT}_n$ [cf. Tab. 2.1]
(Inc)	$[i\,cstit]\varphi \to [Ag\,cstit]\varphi$

Table 2.2: The axiom system $\mathsf{A\text{-}STIT}^{Ag}_n$

2.2.10. THEOREM. [Herzig and Schwarzentruber, 2008] *If $n > 2$, then the problem of deciding satisfiability of a formula of $\mathcal{L}_{\mathsf{A\text{-}STIT}^G_n}$ is undecidable, and the logic $\mathsf{A\text{-}STIT}^G_n$ is not finitely axiomatizable.*

In spite of this result, Schwarzentruber [2012] has shown that, by imposing specific restrictions on the groups that the language can talk about, it is possible to obtain decidable and finitely axiomatizable fragments of group atemporal STIT. In particular, the fragment $\mathsf{A\text{-}STIT}^{Ag}_n$ of $\mathsf{A\text{-}STIT}^G_n$ having $[Ag\,cstit]$ and, for all $i \in Ag$, $[i\,cstit]$ as the only STIT operators is decidable and finitely axiomatizable (an axiomatization is displayed in Table 2.2). But there are other options to circumvent the impossibility of axiomatizing group STIT. Call *complete additivity* the property (characterizing group STIT) that a group action at a moment is the intersection of some individual actions available at that moment, one for each agent in the group. A first way to bypass the negative results from Herzig and Schwarzentruber [2008] is to relax either the notion of group action by dropping the requirement of complete additivity [see, e.g., Balbiani et al., 2008a] or the notion of choice function by dropping the requirement that action sets at a moment partition the histories passing through that moment [see, e.g., Broersen, 2011b]. An alternative option is to assume that the number of choices available to a group at a moment is bounded and add additional machinery to analyze group STIT operators in terms of other, less complex, primitives [see, e.g., Herzig and Lorini, 2010]. The first strategy (relaxing either the notion of group choice or the notion of choice function) is inspired by Coalition Logic [Pauly, 2001, 2002], whereas the second strategy (analyzing STIT operators) is inspired by Propositional Dynamic Logic [Harel et al., 2000]. We will discuss these frameworks in Section 2.3. Anticipating on what is to come, the systems we present in Chapters 3 and 4 are an instance of the second strategy.

2.2.3 Kripke semantics for $\mathcal{L}_{\mathsf{STIT}^{Ag}_n}$

We conclude Section 2.2 by briefly going back to Kripke semantics for STIT, which is at the heart of the completeness proof of the system presented in Chapter 3. In Section 2.2.1, we saw that we can simulate moments and choices in a Kripke frame by using equivalence relations between possible states. To account for the flow of time, Lorini [2013] recently combined Kripke semantics for atemporal STIT with ideas underlying the Ockhamist models proposed by Zanardo

[1996].[10] With respect to the Kripke A-STIT models of Definition 2.2.5, the main novelties are, first, the explicit introduction of the moment-partition (represented by an equivalence relation $R_\Box$ between states) and, second, the introduction of relations R_{G} and R_{H} between states, used to simulate the future and the past on a history. Lorini [2013] also introduces, as an extra element of frames, an equivalence relation R_{Ag} between states representing the choices available to the group of all agents.

2.2.11. Definition (Kripke STIT frame). A Kripke STIT frame is a tuple

$$\langle W, R_\Box, R_{Ag}, R, R_{\mathsf{G}}, R_{\mathsf{H}} \rangle$$

where $W \neq \varnothing$ is a set of possible states, $R_\Box \subseteq W \times W$ and $R_{Ag} \subseteq W \times W$ are equivalence relations, and $R : Ag \to 2^{W \times W}$ assigns to every agent i an equivalence relation R_i on W. Finally, $R_{\mathsf{G}} \subseteq W \times W$ and $R_{\mathsf{H}} \subseteq W \times W$ are, respectively, the *future relation* and the *past relation*. The elements of Kripke STIT frames are required to satisfy the following conditions:

1. *Partition Refinement*: for all $i \in Ag$, $R_i \subseteq R_\Box$.

2. *Independence of Agents*: for all $w_1, \ldots, w_n \in W$, if $w_i \in R_\Box(w_j)$ for all $i, j \in \{1, \ldots, n\}$, then $\bigcap_{i \in Ag} R_i(w_i) \neq \varnothing$.

3. *Complete additivity*: $R_{Ag} = \bigcap_{i \in Ag} R_i$.

4. *No Choice Between Undivided Histories*: for all $w_1, w_2, w_3 \in W$, if $w_1 R_{\mathsf{G}} w_2$ and $w_2 R_\Box w_3$, then there is $v \in W$ s.t. $w_1 R_{Ag} v$ and $v R_{\mathsf{G}} w_3$.

5. *Properties of* R_{G}: for all $w, w_1, w_2, w_3 \in W$,

   Seriality: there is $w' \in W$ s.t. $w R_{\mathsf{G}} w'$.

   Transitivity: if $w_1 R_{\mathsf{G}} w_2$ and $w_2 R_{\mathsf{G}} w_3$, then $w_1 R_{\mathsf{G}} w_3$.

   Future-linearity: if $w_1 R_{\mathsf{G}} w_2$ and $w_1 R_{\mathsf{G}} w_3$, then either $w_2 R_{\mathsf{G}} w_3$ or $w_3 R_{\mathsf{G}} w_2$.

   Strong Irreflexivity: if $w_1 R_\Box w_2$, then it is not the case that $w_1 R_{\mathsf{G}} w_2$.

6. *Properties of* R_{H}: for all $w_1, w_2, w_3 \in W$,

   Converse: $R_{\mathsf{H}} = R_{\mathsf{G}}^{-1} = \{(w, v) \in W \times W \mid v R_{\mathsf{G}} w\}$.

   Past-linearity: if $w_1 R_{\mathsf{H}} w_3$ and $w_2 R_{\mathsf{G}} w_3$, then either $w_1 R_{\mathsf{G}} w_2$ or $w_2 R_{\mathsf{G}} w_3$.

[10] Previous works adding a temporal dimension to Kripke models for atemporal STIT include Herzig and Lorini [2010] and Schwarzentruber [2012]. See Ciuni and Lorini [2018] for a survey of different semantics for temporal STIT and a comparison between them.

The picture represents the Kripke STIT frame corresponding to the STIT frame depicted in Fig. 2.2. As before, time flows upwards. Nine states, w_1 to w_9, are depicted, standing for nine indices. Thick rectangles represent cells in the moment-partition, corresponding, from bottom to top, to moments m_1 and m_2 in Fig. 2.2. The agent-partition for agent 1 coincides with the rows of the grids inside the thick rectangles and the agent-partition for agent 2 with the columns.

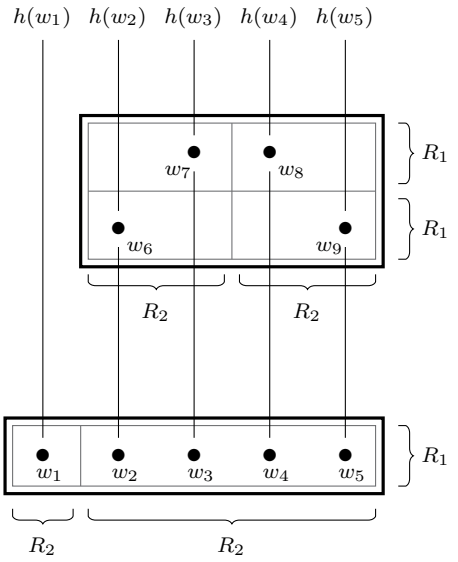


Figure 2.3: A Kripke STIT frame

The first three conditions are familiar: the moment-partition is refined by every choice-partition; group choices are intersections of individual choices; any choice available to any agent at a moment is compatible with any choice available to any other agent at that moment.

The properties of R_{G} and R_{H} ensure that the set $h(w) = R_{\mathsf{H}}(w) \cup \{w\} \cup R_{\mathsf{G}}(w)$ is linearly ordered by R_{G}. Recall that states belonging to the same cell in the moment-partition witness different histories passing through the moment represented by that cell. $h(w)$ is thus the set of witnesses of a history, one for each moment on it – more simply, it is the history witnessed by w.

The last condition, i.e., no choice between undivided histories, is crucial: it ensures, at the same time, that R_{G} induces a tree-like ordering on moments – in particular, that this ordering is linear in the past [see Lorini, 2013, Proposition 3] – and that no choice takes apart undivided histories. To explain, consider states w_2, w_6, and w_7 in the Kripke STIT frame depicted in Figure 2.3. Since $w_2 R_{\mathsf{G}} w_6$, w_2 and w_6 witness the same history, which is one that passes first through moment $R_{\Box}(w_2)$ and then through moment $R_{\Box}(w_6)$. In addition, since $w_6 R_{\Box} w_7$, w_7 witnesses a history that also passes through moment $R_{\Box}(w_6)$. Now, if the history witnessed by w_7 were not witnessed by a state in $R_{\Box}(w_2)$, then the moment represented by $R_{\Box}(w_6)$ would have two different pasts. In addition, if this history were witnessed by a state in $R_{\Box}(w_2)$ but not in $R_{Ag}(w_2)$, then there would be a choice available to some agent at $R_{\Box}(w_2)$ taking apart the histories witnessed by w_6 and w_7, which are undivided at $R_{\Box}(w_2)$ (both histories pass

through $R_{\Box}(w_6)$).

As usual, a *Kripke* STIT *model* is a Kripke STIT frame supplied with a valuation function $\nu : Prop \to 2^W$. The modal formulas of $\mathcal{L}_{\mathsf{STIT}_n^{Ag}}$are then interpreted as follows, where M is a Kripke STIT model and w a state in M:

$$\begin{array}{lll}
M, w \models \mathsf{G}\varphi & \text{iff} & \text{for all } w' \in W, \text{ if } wR_{\mathsf{G}}w', \text{ then } M, w' \models \varphi \\
M, w \models \mathsf{H}\varphi & \text{iff} & \text{for all } w' \in W, \text{ if } wR_{\mathsf{H}}w', \text{ then } M, w' \models \varphi \\
M, w \models \Box\varphi & \text{iff} & \text{for all } w' \in W, \text{ if } wR_{\Box}w', \text{ then } M, w \models \varphi \\
M, w \models [i\, cstit]\varphi & \text{iff} & \text{for all } w' \in W, \text{ if } wR_{i}w', \text{ then } M, w' \models \varphi \\
M, w \models [Ag\, cstit]\varphi & \text{iff} & \text{for all } w' \in W, \text{ if } wR_{Ag}w', \text{ then } M, w' \models \varphi
\end{array}$$

2.2.12. Theorem. [Lorini, 2013] *The axiom system* STIT_n^{Ag}*, defined by the axioms and rules in Table 2.3, is sound and complete with respect to the class of all Kripke STIT frames.*

($\mathsf{A\text{-}STIT}_n^{Ag}$)	Axioms and rules of $\mathsf{A\text{-}STIT}_n$ [cf. Tab. 2.1]		
(UH)	$\mathsf{F}\Diamond\varphi \to \langle Ag\, cstit\rangle \mathsf{F}\varphi$		
($\mathsf{KD4_G}$)	Axioms of $\mathsf{KD4}$ for G	($\mathsf{K_H}$)	Axioms of K for H
($\mathsf{C_{GH}}$)	$\varphi \to \mathsf{GP}\varphi$	($\mathsf{C_{HG}}$)	$\varphi \to \mathsf{HF}\varphi$
($\mathsf{Lin_G}$)	$\mathsf{PF}\varphi \to (\mathsf{P}\varphi \vee \mathsf{F}\varphi)$	($\mathsf{Lin_H}$)	$\mathsf{FP}\varphi \to (\mathsf{P}\varphi \vee \mathsf{F}\varphi)$
($\mathsf{RN_G}$)	From φ infer $\mathsf{G}\varphi$	($\mathsf{RN_H}$)	From φ infer $\mathsf{H}\varphi$
(IRR)	From $(\Box\neg p \wedge \Box(\mathsf{G}p \wedge \mathsf{H}p)) \to \varphi$ infer φ, provided p does not occur in φ		

Table 2.3: The axiom system STIT_n^{Ag}

The axiom system STIT_n^{Ag} is obtained by extending $\mathsf{A\text{-}STIT}_n^{Ag}$ with a standard axiomatization for the modalities G and H, the irreflexivity rule IRR [see Lorini, 2013, p. 315 for comments on this], and the axiom UH, which expresses the condition of no choice between undivided histories. According to the latter axiom, what the group of all agents presently does cannot exclude that what might happen in the future will indeed happen sometime in the future.

2.3 STIT and logics for multiagent systems

Together with the analysis of the mathematical properties of STIT, the last decades have seen an increased interest in the connection between STIT and other logics for multiagent systems (MAS). Among them, we can distinguish logics that, like STIT, allow us to reason about the *agency* of individuals or groups in guaranteeing certain results, and logics that, unlike STIT, allow us to reason

about the *types of action* that guarantee certain results.[11] The main logics of agency in the former group that have been explicitly related to STIT are Coalition Logic (CL) [Pauly, 2001, 2002] and Alternating-time Temporal Logic (ATL) [Alur et al., 2002; Goranko and van Drimmelen, 2006], while the main logic of action in the latter group that has been used in connection with STIT is Propositional Dynamic Logic (PDL) [Harel et al., 2000]. In this section we focus on how STIT relates to CL and PDL, while we mention the connection with ATL only in passing.[12] Besides locating STIT in the panorama of logics for MAS, this will allow us to pinpoint some key ideas that will be important in later chapters.

Preliminary remark. The technical details presented in this section will not be presupposed in what follows. We thus encourage the reader to focus on the main concepts and come back to the details later on, in case they are interested.

2.3.1 Connection of STIT with CL and ATL

CL is a modal logic of agency that was developed by Pauly [2001, 2002] with the aim of formalizing what it means for a group of agents, or coalition, to have the ability to ensure a certain result in a strategic game. Letting Ag and $Prop$ be defined as above, the set $\mathcal{L}_{\mathrm{CL}}$ of formulas of the language of CL is generated by the following grammar:

$$\varphi := p \mid \neg\varphi \mid (\varphi \wedge \varphi) \mid [I]\varphi$$

where $p \in Prop$ and $I \subseteq Ag$. A modal formula like $[I]\varphi$ is read "coalition I has the ability to guarantee φ, no matter what the other agents do" or, for short, "coalition I is *effective* for φ." The semantics for this is given in terms of functions, called *effectivity functions*, that assign to every coalition I and state w the set of propositions for which I is effective at w.[13] Since Pauly is concerned with the notion of effectivity *in strategic games*, effectivity functions are first introduced relative to *game models*, which are essentially extensive form games with simultaneous moves [see Osborne and Rubinstein, 1994]. Game models

[11] We are restricting attention to those logics for MAS that model the strategic structure of the environment, that is, what agents can bring about, either individually or in group. There is another cluster of logics for MAS that aim at modeling the epistemic attitudes of the agents (especially, their beliefs, desires, or intentions) and investigate issues related to their rationality. Among others, epistemic logics [see van Benthem, 2011; van Ditmarsch et al., 2008; Fagin et al., 1995], intention logic [Cohen and Levesque, 1990], and the belief-intention-desire model (BDI) [see Wooldridge, 2000] fall in this area. Concise overviews can be found in Herzig [2015] and van der Hoek and Wooldridge [2012].

[12] For more comparisons among logics for MAS we refer the reader to Goranko and Jamroga [2005] and Herzig [2015].

[13] In the tradition of philosophical logic, by *proposition* we mean a set of possible states, namely those at which the proposition is true. In probability theory and game theory this is often called an *event*.

encode the following intuitive view: At every state w, every agent is endowed with a set of available actions and can choose to perform one action from this set. Taken together, the actions chosen by all agents at w uniquely determine the next state of the world. The formal definition, which we take, with minor changes in notation, from Broersen et al. [2006b],[14] is as follows:

2.3.1. Definition (Game model). Let Ag and $Prop$ be defined as above. A game model is a tuple

$$\Gamma = \langle W, \{A_{i,w} \mid i \in Ag, w \in W\}, o, \nu \rangle$$

where $W \neq \varnothing$ is a set of states, $\nu : Prop \to 2^W$ is a valuation function, and, for every $i \in Ag$ and $w \in W$, $A_{i,w} = \{a_{i,w}, b_{i,w}, c_{i,w}, \dots\}$ is a set of actions available to agent i at w. For every coalition $I \subseteq Ag$ and state $w \in W$, $A_{I,w} = \Pi_{i \in I} A_{i,w}$ is the set of actions available to coalition I at w and $A_{Ag} = \bigcup_{w \in W} A_{Ag,w}$ is the set of *action profiles over* Γ. The last element of Γ is the *outcome function* $o : A_{Ag} \to W$ which assigns to every action profile over Γ a unique outcome state.

As in game theory, in game models individual actions are given an abstract representation: each $A_{i,w}$ is a set of action labels. In addition, as in STIT semantics, the actions available to a coalition at w are modeled by taking together some actions available to its members at w (one for each member). The actions available to the grand coalition Ag (called *action profiles*) have a special role: they determine a transition from a given state to a unique outcome state. The outcome function o keeps track of the transition determined by every action profile.

The outcome function o can be extended to any coalition in the following way. Where $I \subseteq Ag$, let $-I$ be $Ag \setminus I$ and, where $\alpha_{I,w} \in A_{I,w}$, let $\alpha_{I,w}\alpha_{-I,w}$ be the action profile induced by $\alpha_{I,w}$ and $\alpha_{-I,w}$. For every $I \subseteq Ag$, $w \in W$, and $\alpha_{I,w} \in A_{I,w}$, we define:

$$o(\alpha_{I,w}) = \{o(\alpha_{I,w}\alpha_{-I,w}) \mid \alpha_{-I,w} \in A_{-I,w}\}$$

Intuitively, every state in $o(\alpha_{I,w})$ is a possible outcome of $\alpha_{I,w}$ (i.e., a state that might result if coalition I performs $\alpha_{I,w}$). We are now ready to define the notion of effectivity function for a game model.

2.3.2. Definition (Effectivity function for a game model). Where Γ is a game model, an effectivity function for Γ is a map $E^\Gamma : 2^{Ag} \times W \to 2^{2^W}$ defined by setting, for every $I \subseteq Ag$, $w \in W$, and $P \in 2^W$:

$$\begin{aligned} P \in E^\Gamma_I(w) \quad &\text{iff} \quad \text{there is } \alpha_{I,w} \in A_{I,w} \text{ s.t. } o(\alpha_{I,w}) \subseteq P \\ &\text{iff} \quad \text{there is } \alpha_{I,w} \in A_{I,w} \text{ s.t., for all } \alpha_{-I,w} \in A_{-I,w}, \\ &\qquad o(\alpha_{I,w}, \alpha_{-I,w}) \in P \end{aligned}$$

[14] As pointed out by the authors, the differences with Pauly's definition are only "cosmetical."

Whenever $P \in E_I^\Gamma(w)$, we say that I *is effective for* P *at* w.

In words, coalition I is effective for P at w just in case there is an action available to I at w that results in a P-state, no matter what the other agents do (where a P-state is a state included in P). Where Γ is a game model and w a state in Γ, the semantics of formulas like $[I]\varphi$ is then defined as follows:

$$\Gamma, w \models [I]\varphi \text{ iff } \{w' \in W \mid \Gamma, w' \models \varphi\} \in E_I^\Gamma(w) \qquad (\mathsf{Sem}_{[I]})$$

Hence, $[I]\varphi$ is true at w just in case I is effective at w for the proposition expressed by φ. Notice the $\exists\forall$ pattern of quantification hidden in this clause: given Definition 2.3.2, $\mathsf{Sem}_{[I]}$ says that $[I]\varphi$ is true at w just in case *there is* an action $\alpha_{I,w}$ available to I at w such that *all* possible outcomes of $\alpha_{I,w}$ are φ-states. From a conceptual point of view, this suggests that the notion of ability modeled by CL is the same as the one modeled by STIT (but more on this in a moment). From a technical point of view, it reveals that actions are not essential elements of the semantics after all. And, in fact, Pauly [2002] shows that an equivalent semantics can be given by replacing game models with *coalition models*, which are tuples $\langle W, E, \nu\rangle$, where W and ν are as before and $E : 2^{Ag} \times W \to 2^{2^W}$, called a *playable effectivity function*, satisfies the following properties: for every $w \in W$, $P, P' \in 2^W$, and $I, J \subseteq Ag$ such that $I \cap J \neq \varnothing$,

1. *No coalition is effective for the impossible proposition*: $\varnothing \notin E_I(w)$.
2. *Every coalition is effective for the necessary proposition*: $W \in E_I(w)$.
3. *Ag-maximality*: if $P \notin E_{Ag}(w)$, then $W \setminus P \in E_\varnothing(w)$.
4. *Outcome monotonicity*: if $P \subseteq P'$ and $P \in E_I(w)$, then $P' \in E_I(w)$.
5. *Superadditivity*: if $P \in E_I(w)$ and $P' \in E_J(w)$, then $P \cap P' \in E_{I\cup J}(w)$.

The complete axiomatization of CL provided by Pauly [2002] is displayed in Table 2.4. As it is immediate to see, the axioms express in the language the properties of playable effectivity functions. Focusing on the most interesting ones, axiom N, which corresponds to *Ag-maximality*, says that, if the grand coalition cannot guarantee φ, then $\neg\varphi$ will be true in the outcome state, regardless of what any agent does. This reflects that the actions available to the grand coalition determine a unique outcome state. Axiom M expresses the requirement of *outcome monotonicity*, according to which, if a coalition is effective for a proposition, then it must also be effective for the consequences of that proposition. Finally, axiom S corresponds to *superadditivity* and says that disjoint coalitions can jointly achieve whatever they can achieve separately.

Now, CL and STIT seem to share several features, like the strategic interpretation of agency (i.e., the focus on what agents can achieve, *no matter what*

(CPL)	All classical propositional tautologies
($\bot$)	$\neg[I]\bot$
($\top$)	$[I]\top$
(N)	$\neg[Ag]\varphi \rightarrow [\varnothing]\neg\varphi$
(M)	$[I](\varphi \wedge \psi) \rightarrow [I]\varphi$
(S)	$[I]\varphi_1 \wedge [J]\varphi_2 \rightarrow [I \cup J](\varphi_1 \wedge \varphi_2)$ provided that $I \cap J = \varnothing$
(MP)	From $\varphi \rightarrow \psi$ and φ, infer ψ
(RE)	From $\varphi \leftrightarrow \psi$, infer $[I]\varphi \leftrightarrow [J]\psi$

Table 2.4: The axiom system CL

the other agents do) and the analysis of ability. But can the connection be made formally precise? A positive answer was given by Broersen et al. [2006b],[15] who proved that CL can be embedded into STIT, provided that two assumptions are added to standard STIT theory: (1) that time is discrete and without end,[16] and, correspondingly, that the language includes the "next" operator X (a formula like $\mathsf{X}\varphi$ means "φ is true at the next moment on the current history"); (2) that every choice available to the group of all agents at a moment determines the next moment, in the sense that all histories it includes pass through a unique next moment. Let $\mathcal{L}_{\mathsf{XSTIT}^G_n}$ be the set of formulas obtained by extending the set $\mathcal{L}_{\mathsf{STIT}^G_n}$ of formulas of group atemporal STIT with formulas of form $\mathsf{X}\varphi$. The proof that CL can be embedded into STIT is based on the definition of a translation function mapping formulas from $\mathcal{L}_{\mathrm{CL}}$ into formulas from $\mathcal{L}_{\mathsf{XSTIT}^G_n}$. Broersen et al. [2006b] define the translation function $tr : \mathcal{L}_{\mathrm{CL}} \rightarrow \mathcal{L}_{\mathsf{XSTIT}^G_n}$ as follows:

$$tr(p) = \Box p \qquad tr(\varphi \wedge \psi) = tr(\varphi) \wedge tr(\psi)$$
$$tr(\neg\varphi) = \neg tr(\varphi) \qquad tr([I]\varphi) = \Diamond[I\, cstit]\mathsf{X}\varphi$$

Hence, as suggested above, formulas like $[I]\varphi$ essentially express ability in the sense of STIT. The translation also reveals that the modalities of CL can be analyzed as the fusion of three modal operators: one for historic possibility, one for agency, and one for what happens next. Thus, unlike the STIT formula $\Diamond[I\, cstit]\varphi$, the CL formula $[I]\varphi$ is intrinsically "future oriented." This reflects the fact that, in the semantics of CL, the dynamics of actions is not abstracted away: actions determine possible transitions to next outcome states, which do not necessarily coincide with the state at which the agents decide to act.

[15] The authors acknowledge Wölfl [2004] as the first attempt to connect the two frameworks at a conceptual level.

[16] A BT structure $\langle Mom, < \rangle$ is discrete and without end moments when $<$ has the following property: for every $m_1, m \in Mom$, if $m_1 < m$, then there is a moment m_2 such that $m_1 < m_2 \leq m$ and, for no moment m_3, $m_1 < m_3 < m_2$. This property ensures that, for every moment m and history $h \in H_m$, there is a moment on h occurring immediately after (or next to) m.

To give a hint of how the proof presented by Broersen et al. [2006b] goes through, assumptions (1) and (2) ensure that the translations of the axioms of CL (and, in particular, of axiom N) are valid in STIT and that the translated inference rules are truth-preserving. Given Pauly's completeness theorem, this suffices to conclude that the translation of every CL-validity is valid in STIT. The proof of the converse implication (i.e., the translation of every formula satisfiable in a game model is satisfiable in a STIT model) is based on the construction of a STIT model $\mathcal{M}^\Gamma = \langle Mom^\Gamma, <^\Gamma, Ch^\Gamma, \pi^\Gamma \rangle$ from a game model $\Gamma = \langle W, \{A_{i,w} \mid, i \in Ag, w \in W\}, o, \nu \rangle$. We briefly sketch the main steps of the construction to clarify the relation between the two semantics. To keep the presentation simple, we restrict attention to game models where transitions do not give rise to loops.[17] $\mathcal{M}^\Gamma$ is defined as follows:

1. Mom^Γ is just W: states in Γ can be viewed as moments in a STIT model.

2. $<^\Gamma$ is built from the outcome function o. The idea is that every action profile transitions a moment in a unique next moment, and so $w <^\Gamma w'$ holds if there is a sequence of transitions such that the first transition starts at w, the last transition ends at w', and each transition in the sequence starts at the state where the previous transition ends. In other words: time in Γ emerges as a by-product of the actions of the agents.

3. For every agent i and state w, Ch_i^w is built from $A_{i,w}$. In particular, for any action $a_{i,w} \in A_{i,w}$, the set of histories in H_w on which a possible outcome of $a_{i,w}$ occurs is a choice-cell in Ch_i^w. In other words: Ch_i^w is obtained by identifying actions in $A_{i,w}$ with the sets of their possible outcomes.

4. For every index w/h in $\mathcal{M}^\Gamma$ and $p \in Prop$, $w/h \in \pi^\Gamma(p)$ holds just in case $w \in \nu(p)$. This entails that $\Gamma, w \models p$ just in case p is settled true at w in $\mathcal{M}^\Gamma$, which, in turn, explains the translation of propositional variables.

Let us conclude with some brief remarks on the relation between CL, ATL and STIT. While CL is a framework to reason about the control a coalition can exercise on what happens *next* by choosing an *action*, ATL [Alur et al., 2002; Goranko and van Drimmelen, 2006] is a framework to reason about the control a coalition can exercise on what happens *in the long run* by choosing a *strategy*. The key modal formulas of ATL have form $\langle\!\langle I \rangle\!\rangle \varphi$, read "coalition I has a strategy that guarantees that φ." Goranko [2001] showed that CL is a fragment of ATL, by identifying the CL formula $[I]\varphi$ with the ATL formula $\langle\!\langle I \rangle\!\rangle \mathsf{X}\varphi$. The question then naturally arises whether the embedding of CL into STIT can be extended to the whole ATL. A positive answer was again given by Broersen et al. [2006a], who proved that ATL can be embedded into the extension of STIT with strategies, called *strategic* STIT, proposed by Horty [2001, Chapter 7] (as before, under the

[17] If they do, Γ needs first to be unraveled [see Blackburn et al., 2001, p. 63].

proviso that some additional assumptions are granted). In particular, it turns out that the ATL formula $\langle\!\langle I \rangle\!\rangle \mathsf{X}\varphi$ can be identified with the strategic STIT formula $\Diamond_s[I\ scstit]\mathsf{X}\varphi$, where $\Diamond_s[I\ scstit]$ is a fused operator of long-term strategic ability. This result shows that STIT logics are the most general logics of agency on the market. Importantly, besides allowing us to reason about what individual agents and groups *can* do, STIT, unlike CL and ATL, also allows us to reason about what they *actually* do. As we will see below, this is also one of the features that distinguishes STIT from PDL.

2.3.2 Comparison between STIT and PDL

So far, we have considered frameworks allowing us to reason about the agency of individual agents or groups in bringing about certain effects. But what about the performed actions themselves? As we mentioned in Section 2.1, STIT semantics represents actions as *action tokens*, that is, as concrete, particular events occurring at a unique space-time location, while there is nothing in the semantics that allows us to group together actions of the same *type* [cf. Horty and Pacuit, 2017, p. 617]. So, we can represent, e.g., Ann's particular dropping of a particular glass in a particular situation and say that, by performing this particular action, Ann ensures that the particular glass will break. But we cannot classify Ann's particular action as an action of type "dropping a glass" and say, for instance, that Ann broke the glass by dropping it rather than, say, by hitting it against the table, or that she performs this type of action twice in a row, or that Bob just did the same type of thing (and, yet, he did not break a glass). Game models, unlike STIT models, include action labels that can be interpreted as action types. Yet, the object language of CL does not include any explicit reference to actions – which is why, ultimately, actions are inessential elements of the semantics.

Unlike STIT and CL, PDL [Fischer and Ladner, 1979; Harel, 1984; Harel et al., 2000], the propositional counterpart of Pratt's [1976] Dynamic Logic (DL), is a modal logic of programs that was specifically designed to reason about the possible executions of different *types* of programs. Although it originated in computer science, PDL was later applied to deontic logic [Meyer, 1988] and philosophy of action [Segerberg, 1992] on the basis of the analogy between computer programs and actions: as Meyer [1988, p. 110] has it, "[o]ne has to realize that a computer program is in fact nothing but a sequence of actions of a certain kind." Following this insight, from now on, we will take PDL as a logic of action.

The characterizing feature of PDL is that its language contains two categories of expressions: (names of) action types and formulas. The modal formulas of the logic have form $[\alpha]\varphi$, where α is (a name of) an action type, and are read "after every possible execution of α, φ is true" or "doing α necessarily results in a state in which φ is true." The action type α can either belong to a given set Atm of atomic types or result from more basic types, using either one of the following

operations:[18]

1. *Sequential composition*: $\alpha;\beta$ is the type that is instantiated whenever β is performed after α;

2. *Nondeterministic composition*: $\alpha \cup \beta$ is the type that is instantiated whenever either α or β is performed;

3. *Finite iteration*: α^* is the type that is instantiated whenever α is performed repeatedly, for a finite number of times.

The set *Types* of action types is built from *Atm via* the three operations above.

In the semantics, action types are interpreted as binary relations between states. The idea is that actions are *dynamic* entities: as the occurrence of an event, the performance of an action typically results in a change of the world. When we, for instance, switch the light on, the world changes from a state in which the light is off to a state in which it is on; when Ann drops her glass, the world changes from a state in which the glass is in Ann's hand to a state in which the glass is on the floor; and so on. A possible performance, or execution, of an action type can then be thought of as a transition between two possible states: an *initial-state* at which the action starts and an *end-state* at which the action is concluded. By assuming an extensional view on types, this leads to model an action type as a set of transitions between possible states, namely those that correspond to its executions.

2.3.3. Definition (PDL model). Let *Atm* and *Prop* be defined as above. A PDL model is a tuple $\langle W, R, \nu\rangle$, where $W \neq \varnothing$ is a set of possible states, $\nu : Prop \to 2^W$ is a valuation function, and $R : Atm \to 2^{W\times W}$ assigns to every atomic type a a binary relation R_a over W.

Intuitively, wR_aw' means that there is a possible execution of the atomic action a at w that results in state w'. So, $R_a(w) = \{w' \in W \mid wR_aw'\}$ is the set of *possible outcomes of a at w*. The function R is naturally extended to the set of all action types by setting, for every type α and β:

$$R_{\alpha;\beta} = R_\alpha \circ R_\beta \qquad R_{\alpha\cup\beta} = R_\alpha \cup R_\beta \qquad R_{\alpha^*} = R^*_\alpha$$

where $\circ$ takes two binary relations and returns their composition and * takes a binary relation and returns its reflexive transitive closure. The semantics of formulas like $[\alpha]\varphi$ is defined by the following clause,[19] where M is a PDL model and w a state in M:

$$M, w \models [\alpha]\varphi \text{ iff, for all } w' \in W, \text{ if } wR_\alpha w' \text{ then } M, w' \models \varphi \qquad (\mathsf{Sem}_{[\alpha]})$$

[18] More operations can be added, like the *test* operation ? that, for every formula φ, returns the program $?\varphi$, corresponding to "proceed if φ is true; fail otherwise." We here restrict attention to the operations that are the most common in action logic.

[19] The action modalities $[\alpha]$ are the only modalities of PDL.

(CPL)	All tautologies of CPL	(Seq)	$[\alpha;\beta]\varphi \leftrightarrow [\alpha][\beta]\varphi$
($\mathsf{K}_{[\alpha]}$)	The axiom schema of K for $[\alpha]$	(Com)	$[\alpha \cup \beta]\varphi \leftrightarrow [\alpha]\varphi \wedge [\beta]\varphi$
(MP)	From $\varphi \rightarrow \psi$ and φ, infer ψ	(FP)	$[\alpha^*]\varphi \leftrightarrow (\varphi \wedge [\alpha][\alpha^*]\varphi)$
($\mathsf{RN}_{[\alpha]}$)	From φ, infer $[\alpha]\varphi$	(Ind)	$\varphi \wedge [\alpha^*](\varphi \rightarrow [\alpha]\varphi)) \rightarrow [\alpha^*]\varphi$

Table 2.5: The axiom system PDL

So, $[\alpha]\varphi$ is true at w just in case all possible executions of α at w result in a state where φ is true. A well-known axiomatization of PDL is displayed in Table 2.5.[20] We only point out that axioms Seq and Com reveal that, in the star-free fragment of PDL, all operators for complex modalities can be eliminated.

From a conceptual point of view, there are various points of comparison between STIT and CL on the one hand and PDL on the other hand, some of which we have already touched upon. Let us highlight the main issues.

Actions and time. PDL models are clearly more similar to game models than to STIT models: both PDL models and game models are so-called labeled transition systems, that is, structures that represent possible transitions between states, tagged with labels from a given set (in the present case, with the actions that, intuitively, bring them about). Unlike STIT frames, PDL models and game models are *action-driven* rather than time-driven: although it is not explicitly represented, time can be viewed as emerging as a by-product of the performance of actions. In particular, like actions in game models, atomic action types in PDL – possibly combined via the operation $\cup$ of nondeterministic composition – can be viewed as *one-step actions*, determining transitions to a possible next state. By contrast, actions built by using the operations ; of sequential composition and $*$ of finite iteration can be viewed as *courses of action*, determining transitions consisting of multiple steps and possibly leading to states that are far in the future. As in the case of CL, this dynamic conception of actions makes PDL modalities intrinsically "future oriented": the semantics of formulas like $[\alpha]\varphi$ gives them the intended meaning that, *after* the execution of α, φ is true.

Interactions between agents. With respect to STIT and CL, an eye catching feature of PDL is that there is no reference to agents either in the semantics or in the object language. The standard way to deal with this issue [see, e.g., Herzig and Longin, 2004; Lorini and Herzig, 2008; Meyer et al., 1999; Wieringa and Meyer, 1993] is to build complex action types from a set of pairs (a, i), where $a \in Atm$ is an atomic type and $i \in Ag$ is an agent (we write a_i rather than (a, i)

[20] The first axiomatization of PDL was presented by Segerberg [1977], and the first completeness proof was due to Parikh [1978]. A textbook presentation of the proof can be found in Blackburn et al. [2001, Chapter 4.8].

in what follows). Intuitively, a_i is the type of action that is instantiated when agent i performs an action of type a. Let a *modified PDL model* be a PDL model in which the domain of the function R is the set of all pairs a_i included in the language. Then, for any states w, w' in a modified PDL model, $wR_{a_i}w'$ represents the fact that agent i contributes to the transition from w to w' by doing an action of type a. In this way, PDL, like STIT and CL, can be used to reason about the interaction between different agents. In this regard, observe that, unless further assumptions are made, the following are not excluded:

1. An agent i can perform more than one type of action at a time: it is possible that $R_{a_i}(w) \cap R_{b_i}(w) \neq \varnothing$ for some $a \neq b$.
2. An agent j cannot contribute to a possible transition: it is possible that $w' \in R_{a_i}(w)$, for some $a \in Atm$ and $i \in Ag$, but there is no $b \in Atm$ s.t. $w' \in R_{b_j}(w)$.
3. Different agents i and j can have incompatible types of action available to them: it is possible that $R_{a_i}(w) \neq \varnothing \neq R_{b_j}(w)$ but $R_{a_i}(w) \cap R_{b_j}(w) = \varnothing$.

Hence, the action types available to an agent at a state w in a modified PDL model might not partition the set of transitions starting at w. In addition, they might not be independent of the action types available to other agents at w. This means that, unlike in STIT, an agent can prevent another agents from doing certain actions.

Agency and ability. Having introduced agents, we can now consider how agency and ability can be expressed in PDL. Formulas like $[a_i]\varphi$ are interpreted by adapting clause $\mathsf{Sem}_{[\alpha]}$: where w is a state in a modified PDL model M,

$$M, w \models [a_i]\varphi \text{ iff, for all } w' \in W, \text{ if } wR_{a_i}w' \text{ then } M, w' \models \varphi \qquad (\mathsf{Sem}_{[a_i]})$$

Accordingly, $[a_i]\varphi$ is true at a state w in a model M just in case, if agent i performs an action of type a at w, then φ will be true. Notice the conditional form of this reading: unlike the STIT formula $[i\,cstit]\varphi$, $[a_i]\varphi$ does not say that i *actually* brings about φ, but that i brings about φ *if she performs an action of type* a. In this sense, borrowing an expression from Herzig et al. [2018], PDL, like CL, is about *potential agency* and not about *actual agency*.

On the other hand, unlike STIT and CL, PDL is not only about the effects an agent can realize by acting, but also about the types of action she has to do to bring the effects about. This adds an interesting component to the analysis of ability. Consider the following formula, where $\langle a_i \rangle$ abbreviates $\neg[a_i]\neg$:

$$\langle a_i \rangle \top \wedge [a_i]\varphi$$

As it is easy to check, the first conjunct is true at a state w just in case $R_{a_i}(w) \neq \varnothing$. This means that agent i can perform an action of type a at w, that is, this type of action is available to her at w. In addition, according to the second conjunct,

a is a type of action that, if performed by i, ensures that φ is true. But then $\langle a_i \rangle \top \wedge [a_i]\varphi$ means that agent i has the *ability* to bring about φ *by doing* a_i. The familiar $\exists\forall$ pattern can be recovered by quantifying over a_i: letting A_i be the set of action types associated with agent i (and assuming that this set is finite), the formula

$$\bigvee_{a_i \in A_i} (\langle a_i \rangle \top \wedge [a_i]\varphi)$$

expresses that there is an action available to agent i that guarantees the truth of φ, i.e., that i has the ability, in the STIT or CL sense, to bring about φ.

As first highlighted, in the STIT literature, by Herzig and Troquard [2006], the possibility to refer to the types of action by means of which an agent can bring about a certain result is crucial to model an important sense of ability, namely "knowing how."[21] More generally, as already emphasized by Segerberg [1992] and, more recently, by van Benthem and Pacuit [2014], merging STIT and PDL seems to be a necessary step towards a unified and comprehensive framework to reason about agency. In the last decades, several authors, including Broersen [2014b], Herzig and Lorini [2010], Horty [2019], Horty and Pacuit [2017], Segerberg [2002], Troquard and Vieu [2006], Xu [2010, 2012] among others, have taken up this challenge. In the coming chapters in Part I of the dissertation, we will see that supplementing STIT with action types is also key to applying the theory to the analysis of causal responsibility and, relatedly, to the study of counterfactual reasoning concerning what can be, or could have been, done in the course of time. The importance of action types in relation to normative reasoning will emerge in Part II.

[21] For a recent discussion see Horty and Pacuit [2017] and Broersen and Ramírez Abarca [2018].

Part One

Agency and counterfactuals

Chapter 3

Causal responsibility: A first refinement of STIT

Causal responsibility is the kind of responsibility that derives exclusively from the fact that an agent brought about a certain state of affairs, no matter her intentions or beliefs. It is one of the most basic forms of responsibility both in moral and legal reasoning: Moral blameworthiness as well as criminal and civil offenses are usually determined on the basis of two elements, namely the subject's acts (called *actus reus*, or "guilty act," element in the law) and the subject's intentional states (called *mens rea*, or "guilty mind," element in the law). The question whether the subject's acts were guilty precedes the question whether her mind was guilty. First, one determines if what the agent did *actually caused* a negative result, that is, whether the agent can be said to be *causally responsible* for it. *If so*, one further considers *why* the agent did what she did, that is, whether she acted recklessly, knowingly, or, even worse, intentionally.[1]

Given this picture, a central prerequisite for a logic to analyze responsibility attributions is that it supports reasoning about the states of affairs that are *actually* brought about. As we saw in Chapter 2, STIT logic, unlike other logics for MAS, satisfies this prerequisite. It is then not surprising that in the last years STIT has become one of the main reference tools, among logicians, to formally analyze complex notions of moral and legal responsibility.[2] Assuming that formulas such as $[i\,cstit]\varphi$ and $[i\,dstit]\varphi$ (or variants thereof) express that agent i is causally

[1] Moral blameworthiness and the kind of legal responsibility considered here fall in the category of so-called *backward looking responsibility*, i.e., responsibility for what has happened in the past. A different category that we will not consider in what follows is *forward looking responsibility*, i.e., responsibility for what one ought to realize in the future. An example of backward looking responsibility is responsibility for a car accident. An example of forward looking responsibility is responsibility for submitting a review by the deadline. A recent taxonomy of different kinds of responsibility with an analysis of how they relate to one another can be found in Van De Poel [2015].

[2] Recent works in this direction include Broersen [2011a,b, 2014a], Ciuni and Mastop [2009], Duijf [2018], Mastop [2010], Lorini and Schwarzentruber [2011], and Lorini et al. [2014].

responsible for φ, this line of research has primarily focused on studying various aspects of the notion of *mens rea* by relying on extensions of STIT with epistemic operators [see, e.g., Broersen, 2011b; Broersen and Ramírez Abarca, 2018; Herzig and Troquard, 2006; Horty and Pacuit, 2017; Lorini et al., 2014], probabilistic belief operators [Broersen, 2013, 2014a], operators for intentions [Broersen, 2011a], and deontic operators [see Xu, 2015] (to name a few).

In this chapter we take a step back with respect to this trend. Our aim is to bring to the foreground the problem of analyzing ascriptions of *causal* responsibility in STIT and to refine the standard semantics in order to make it suitable to address this problem. Our contribution is to introduce genuinely causal notions in STIT – an issue that, to our knowledge, has been explicitly addressed only by Xu [1997] so far – and to provide a formalization of different, but intimately related, notions of causal responsibility. We show how the interplay between these notions accounts for important aspects in the ascription of individual as well as group responsibility that cannot be detected in the standard semantics.

Outline. We start, in Section 3.1, by clarifying why standard STIT semantics needs to be refined and by presenting a three-phase view on the attribution of causal responsibility. This view motivates us to move to a richer framework, which is presented in Section 3.2. We first supplement STIT frames with action types and with a relation of opposing between action types [Section 3.2.1]. We then introduce the syntax and semantic of our STIT logic with action types and opposing (ALO_n) and present a sound and complete axiomatization [Section 3.2.2]. In Section 3.3 we formalize the main phases in the attribution of causal responsibility and introduce corresponding responsibility operators. After comparing the new operators with the deliberative STIT operator, in Section 3.4 we analyze a number of paradigmatic examples involving both individual and group responsibility. Section 3.5 summarizes the main results and highlights directions for future work. The proofs of completeness and of some key propositions from Section 3.3 are found in Appendix A.

This chapter is based on Baltag et al. [Forthcoming].

3.1 Refining STIT: why and how

In the literature on STIT, formulas like $[i\,cstit]\varphi$ or $[i\,dstit]\varphi$ are typically taken to express that agent i brings about φ, in the sense that the action that agent i performs *causes* φ. This is why standard STIT operators are often used to model, at least to a first approximation, causal responsibility. But there is an important mismatch between the intuitive notion of bringing about underlying ascriptions of causal responsibility and the notion of bringing about modeled in STIT. On the one hand, responsibility for a result presupposes that the result is caused by the agent, where *causing the result is compatible with the possibility that other*

Alice is agent 1 and can either stand still (K_1) or shoot Dan (K_2); Beth is agent 2 and can either stand still (K_3) or hit Alice's arm (K_4). φ stands for "Dan is dead some time in the future" and is only true at index m_1/h_1. h_1 (the thick line) is the actual history.

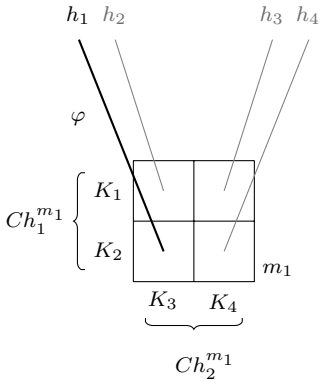


Figure 3.1: Example 3.1.1 in standard STIT semantics

agents could have prevented it. On the other hand, $[i\,cstit]\varphi$ is true when the result (i.e., φ) is guaranteed by the agent, where *guaranteeing the result excludes the possibility that other agents could have prevented it.* This is a straightforward consequence of the condition of independence of agents [condition 2 in Definition 2.1.5]. The latter condition ensures the validity of formulas like $[i\,cstit]\varphi \rightarrow \neg\Diamond[j\,cstit]\neg\varphi$, which say that an agent brings about only those things that no other agent has the ability to block.

The aforementioned mismatch has important consequences on the possibility of using standard STIT semantics to represent causal responsibility (and more elaborate notions) in complex multiagent scenarios. To better appreciate this point, consider the following example.

3.1.1. Example. Alice shoots Dan dead. Beth, a bystander, could prevent Dan's death by hitting Alice's arm, but she remains still, petrified with fear.

Does Alice see to it that Dan is dead in this scenario? Is she causally responsible for it? Figure 3.1 depicts the simplest STIT model of Example 3.1.1, where history h_1 (the thick line) represents the actual course of events. Given this model, the answer to our questions is negative: because of what Beth can do, Alice's action does not guarantee Dan's death; hence, Alice does not see to it that Dan is dead. Yet, a jury assessing the case would certainly disagree: after all, Alice did pull the trigger and cause Dan's death![3]

[3]This consequence of the condition of independence of agents is well-known in the literature, see Royakkers and Hughes [2020, p. 335] for a recent example. There are two potential reactions to this problem: first, to grant that STIT only applies to situations in which agents act independently; second, to argue that cases like Example 3.1.1 require a more elaborate representation, e.g., one in which Alice has a choice corresponding to a way of shooting that guarantees Dan's death. We agree with the former reaction and view our proposal as a way to extend the range of applications of STIT.

In this chapter, we refine STIT in order to make it suitable to model causal agency and analyze causal responsibility in basic multiagent scenarios of this sort. To do this, we will view the attribution of causal responsibility for φ as a task involving the following three preliminary phases:

1. selection of the relevant context;

2. identification of the potential causes of φ;

3. selection of the actual causes of φ.

Similar phases characterize causal analyses in terms of causal models *á la* Halpern and Pearl [Pearl, 2000; Halpern and Pearl, 2005; Halpern, 2016], where the identification of the actual causes of an event presupposes the selection of a set of variables describing the relevant situation and the specification of a set of structural equations describing the assumed causal influences between the variables.[4] Causal analyses in the law and in everyday practice also follow a similar pattern.[5] Here, we will understand the three phases in terms of the following three corresponding questions:

1. *Who was involved and what actions did they do?*

2. *Who did something that was expected to result in φ?*

3. *Who, in addition, did something that actually contributed to φ?*

We devote the rest of this section to clarify the content of each phase and explain how we will enrich STIT semantics in order to account for it.

[4]In a nutshell, a causal model is a tuple $\langle \mathcal{U}, \mathcal{V}, \mathcal{R}, \mathcal{F} \rangle$, where: $\mathcal{U}$ is a set of *exogenous variables* (variables whose values are determined by factors that are not modeled); $\mathcal{V}$ is a set of *endogenous variables* (variables whose values are determined by factors that are modeled); $\mathcal{R}$ is a function assigning to every variable in $\mathcal{U} \cup \mathcal{V}$ a range of possible values; $\mathcal{F}$ is a function assigning to every variable in $\mathcal{V}$ a structural equation. To illustrate, suppose that a house in the middle of a forest burns down and we want to establish what the cause was. We know that there were two arsonists, Ann and Bob, in the forest and we are confident that one of them started a fire that destroyed the house. We can build a causal model of the situation as follows. $\mathcal{U}$ includes two variables: U_A for Ann's psychological conditions, which takes value 1 if such conditions leads Ann to drop a lit match; U_B for Bob's psychological conditions, which takes values 1 and 0 in the same way as U_A. $\mathcal{V}$ includes three variables: A for Ann, which takes value 1 if Ann drops a lit match and 0 if not; B for Bob, which takes values 1 and 0 in the same way as A; H for the house, which takes value 1 if the house burns down and 0 if not. Finally, $\mathcal{F}$ includes the following equations: $A = U_A$, $B = U_B$, $H = min(A, B)$. According to the equations, whether Ann (resp. Bob) drops a lit match depends on Ann's (resp. Bob's) psychological conditions. In addition, whether the house burns downs depends on whether either one of the two agents drops a lit match. See Halpern [2016, Chapter 2.1] for more details.

[5]See, e.g., Lagnado and Gerstenberg [2017], where the causal model approach is related to legal and everyday causal reasoning.

3.1.1 Agents and action types

Imagine that we are members of a jury assessing Example 3.1.1. Given the evidence presented in court, our first task is to figure out what happened: Who was involved? What did they do? And, how did they interact? By answering these questions, we select the relevant agents and the relevant actions they were doing when the homicide took place. For instance, we may say that, when Dan died, Alice and Beth were present at the scene, Alice fired a shot, and Beth was standing close to her. In this way, we select Alice and Beth as the relevant agents and the actions of Alice's shooting and Beth's standing as the relevant actions. Agents and actions that we do not deem relevant for our inquiry, like Carl's entering a shop two streets away, David's reading the newspaper in the park, or Beth's smoking a cigarette (while standing), are simply ignored.

The selected agents and actions form what we will call the *relevant context*. The relevant context is a basic description of what happened and it sets the boundaries of our investigation: agents and actions that are left aside cannot be used to draw any conclusion in the causal analysis. In this sense, the selection of the relevant context is similar to the process of choosing the variables constituting a causal model [Pearl, 2000; Halpern and Pearl, 2005; Halpern, 2016]: like the variables of a causal model, the elements of the relevant context determine the language used to frame the situation. Of course, there may be disagreement on which language better serves this purpose. This reflects a basic feature of both legal and everyday causal reasoning. As Halpern [2015, pp. 91-92] has it,

> [E]ven if there is agreement regarding the definition of causality, there may be disagreement about which model better describes the real world. I view this as a feature of the definition. It moves the question of actual causality to the right arena – debating which of two (or more) models of the world is a better representation of those aspects of the world that one wishes to capture and reason about. This, indeed, is the type of debate that goes on in informal (and legal) arguments all the time.

As causal models are silent on what counts as an appropriate choice of the variables, our framework will be silent on what counts as an appropriate choice of the relevant context.[6] In addition, the selection of the relevant context itself will be implicit in our modeling practice, that is, in the selection of one of our models rather than another to represent a given scenario. Even so, we will describe the given scenario in terms of both *relevant agents* and *relevant actions*, where

[6] Assuming this view may raise the worry that it gives the modeler too much flexibility, and that some general criteria to be followed in the modeling practice should be laid down. We leave a more in-depth discussion of this issue to another venue. In the literature on causal models, it has been addressed by Halpern and Hitchcock [2010] among others. See Halpern [2016, Chapter 4] for more references.

we think of the latter as obtained by abstracting away from the features of the concrete conduct of the agents that are immaterial for the inquiry. To do this, we will supply STIT with *action types*.

We assume a minimal view on action types: Unlike action tokens, they are abstract and repeatable events. In addition, the degree of granularity of the action types used to describe a given situation depends on the features that are relevant to the modeler. Since *who* performs an action is always relevant to attributing responsibility, we will associate action types with their authors. Accordingly, instead of working with overly generic action types like *shooting* and *standing*, we will work with *individual action types* like *Alice's shooting* and *Beth's standing* and with *group action types* like *Alice's shooting and (simultaneously) Beth's standing*.[7] Again, depending on what is relevant to the analysis, action types can be more specific than this. For instance, the action types *Alice's shooting Dan*, *Alice's shooting Dan with a rifle*, *Alice's shooting Dan with a rifle while stepping back and yelling at Beth* can all be used to describe what happened.[8]

In the following, we will restrict our attention to situations in which the relevant agents act simultaneously and all actions are completed in *one step*. In line with PDL [Harel et al., 2000], we will think of the performance of an action type at a moment as determining a set of transitions to possible next moments. This will lead us to introduce STIT operators expressing that the choices made by an individual agent or a group at a moment have effects in the next moment, in the spirit of XSTIT [Broersen, 2009, 2011b]. In addition, we will model events that occurred in the past as if they were occurring at the present moment. So, in the case of Example 3.1.1, our language will allow us to say that Alice *is shooting* Dan and Dan *will be dead* in the next moment, rather than that Alice *shot* Dan and Dan *is dead* now. Although not essential to our proposal, this shift of perspective will simplify the definitions of the main notions.

3.1.2 Expected results and opposing relation

As members of a jury, once we have selected the relevant context (involved agents and their actions), our next task is to identify the potential causes of the negative fact before us (call it φ). We typically do this by considering whether some of the involved agents did something, by themselves or with others, that was expected to result in φ. Let us consider an elaborated version of Example 3.1.1.

[7] As we will make precise below, group action types are essentially *conjunctions* of individual action types: they are instantiated when all individual types constituting them are jointly instantiated. Group action types are thus the most elementary kind of group actions: they require neither a cohesive group nor coordinated actions.

[8] Although the types in these (and later) examples are quite specific, note that they are still types: a situation in which Alice shoots Dan with a rifle while stepping back and yelling at Beth is a type of situation that can be realized in different ways and occur in different hypothetical circumstances.

3.1.2. Example. Alice and Carl fire at Dan, aiming at the heart. Diana also fires at Dan, but aims at the legs. Beth remains still, petrified with fear. Dan dies.

In assessing the case, we will agree that Alice and Carl did something that was expected to result in Dan's death (they both aimed at the heart!), while Beth and Diana did not. We will thus include Alice's and Carl's actions in the list of the potential causes of Dan's death, and exclude Beth's and Diana's actions.

But how do we determine the expected results of an action? Let us start from the intuitive idea that, when an agent is doing something, she may not finish what she is doing because of the external interference of some other agents. This happens all the time in everyday life: my flatmate is coming upstairs and I am going downstairs, through a narrow staircase, at the same time; my sister is watering the plants and her partner trips over the garden hose; I am getting off the train when a group of kids decide that they are not going to wait and start boarding; and so on. A simple, natural way to determine what the expected results of an action are in a given situation is to consider what would happen in that situation if all such external interference were removed. If a state of affairs φ obtains in all hypothetical scenarios that are exactly like the current one except that nothing interferes with what agent i is doing, then we would take what agent i is doing to be at least a candidate, potential cause of φ.[9]

In order to capture this view, we will enrich STIT semantics with a *primitive relation of opposing* between individual and group action types. Intuitively, an action type like *Ilaria's going downstairs* opposes another action type like *Aybüke's going upstairs* when the performance of the former is generally assumed to interfere with the performance of the latter.[10] In modeling concrete cases, we establish which individual or group action types oppose a given individual or group action type by relying on our common-sense understanding of the physical and social world and of the meaning of the descriptions we use to represent the cases in question. Of course, this means that there may be disagreement on the extension of the opposing relation. Again, as in the case of causal models, this is not a limit but a feature of our framework: it separates the task of building a causal story of what happened from the task of justifying this story.[11]

[9] This resembles the dynamics of an intervention in causal models [see Woodward, 2003, Chapter 3.1; Halpern, 2016, Chapter 2.1]. When we intervene on the value of a variable in a causal model, all causal dependencies of that variable on other variables are broken. In a similar way, when we determine the expected results of an action, all external interferences with that action are eliminated.

[10] Notice that it is possible to interfere with the performance of an action without blocking it: when Aybüke and I take the stairs in opposite directions but I end up walking back to let her pass, I interfere with Aybüke's action without blocking it.

[11] An interesting question is whether the notion of external interference (and so the relation of opposing) can be further analyzed. An intuitive idea is that, when we describe the action an agent is doing in general terms, our description comes with a goal that we attribute to the agent

We will make three assumptions concerning the opposing relation. First, whether an individual or group action type opposes another individual or group action type does not depend on the concrete situation in which those types are performed: *Alice's shooting Dan* opposes *Dan's skipping rope* whether or not there is a wall between Alice and Dan. In other words, we view opposing as an intrinsic relation between action types. This is a simplifying assumption that can be relaxed without particular difficulties. Second, no individual or group action type opposes itself: the performance of an action does not generate any external interference on the action itself. In this regard, let us emphasize that we take opposing to be a relation between action types associated with their possible authors, not between action types in general. So, the fact that, e.g., the action type *shooting* appears in both *Alice's shooting Dan* and *Dan's shooting Alice* has no implication whatsoever on the possibility that the latter types oppose each other without opposing themselves.[12] Finally, when an individual or group action type A opposes another individual or group action type B, A also opposes any group action type that includes B. The intuitive justification of this assumption is that, by interfering with what someone is doing, we automatically interfere with what that person is doing together with other people. So, by opposing *Aybüke's going upstairs*, *Ilaria's going downstairs ipso facto* opposes *Aybüke's going upstairs and (simultaneously) Tom's ringing the doorbell*.

We do not assume that, when A opposes B, B is also opposed by any group action that includes A, because the latter group action may include elements that "oppose the opposer," so to speak. For instance, *Alice's shooting Dan* interferes with *Dan's skipping rope*, but in the group action *Alice's shooting Dan and Beth's hitting Alice's arm* this interference is canceled by Beth's action. So, even if it includes *Alice's shooting Dan*, the group action *Alice's shooting Dan and Beth's hitting Alice's arm* does not seem to oppose *Dan's skipping rope*. The same example shows that the relation of opposing is not symmetric: *Alice's shooting Dan* opposes *Dan's skipping rope*, but not vice versa.

3.1.3 *But-for* and NESS tests

Let us go back to Example 3.1.2. After including Alice's action and Carl's action in the list of potential causes of Dan's death, we have to decide which of them *actually* caused it: Was it Alice's shot, Carl's shot, or both? The literature

insofar as he is doing that action. For instance, the description *Alice's shooting Dan* comes with the goal *Dan is seriously injured*, or even *Dan is dead*. The notion of external interferences may then be analyzed in terms of incompatibility of goals or probabilistic dependence between goals (e.g., an action interferes with another when the realization of the goal of the former decreases the probability of the realization of the goal of the latter). We leave a full exploration of these possibilities to future work. We thank Valentin Goranko, John Horty, Frederik Van De Putte, and Olivier Roy for insightful discussions on this issue.

[12] We thank an anonymous referee of the book series Trends in Logic for suggesting this example to us.

on actual causality is extensive and we will attempt neither to survey it nor to defend one particular theory over another.[13] Rather, we will simply assume the view adopted in legal theory (especially in tort and criminal law), which is the conception of a cause as a *difference maker* that goes back to Hume [1748] and has been revived by Lewis [1973b, 1986]. This view is condensed in the so-called *but-for* test: what agent i did was an actual cause of a state of affairs φ if, *but for* i's action, φ would not have occurred.[14]

The *but-for* test is appealing for its simplicity and, in many cases, it provides us with intuitive results. Still, it is well-known to fail in cases of overdetermination: in Example 3.1.2, if Alice's shot and Carl's shot were separately sufficient for Dan's death, then neither of them would satisfy the *but-for* test and thus there would be no actual cause of the death. In the legal literature, this limitation has been addressed by using the more flexible NESS test [Wright, 1988, 2013].[15] According to this account, what an agent did was an actual cause of a state of affairs φ if it was part of a minimal sufficient condition for φ that occurred, where a sufficient condition for φ is minimal when none of its parts is a sufficient condition for φ (we will come back to this later).

In order to represent the last preliminary phase in the attribution of causal responsibility, we will implement both the *but-for* and the NESS tests in our framework. We will then use the notion of potential cause, understood in terms of expected results, and the notion of actual cause, understood in terms of *but-for* or NESS conditions, to study corresponding notions of causal responsibility, which we will call *potential*, *strong*, and *plain causal responsibility*.

Before proceeding, we should mention some connections with the existing literature:

1. von Wright [1971] already analyzes necessary, sufficient, and NESS conditions in a branching time framework. Yet, no explicit representation of actions is provided. In addition, the analysis is only carried out at the semantics level.

2. Xu [1997] models a notion of causality inspired by the NESS account in a branching time framework supplied with "events," which resemble action tokens in standard STIT semantics (unlike action tokens, events are not per-

[13] See Mumford and Anjum [2013] for a concise introduction and Beebee et al. [2010] for an extensive survey covering the history of causation, the standard approaches, and the most contemporary developments.

[14] The classic reference on causation in the law is Hart and Honoré [1959]. More recent discussions can be found in Moore [2009, 2019].

[15] The acronym "NESS" stands for "Necessary Element of a Sufficient Set." In philosophy, the NESS test is closely related to Mackie's [1965; 1974] INUS account: an INUS condition is "an *insufficient* but *necessary* part of a condition which is itself *unnecessary* but *sufficient* for the result" [Mackie, 1965, p. 34]. In legal theory, a version of the NESS test appears in Hart's and Honoré's [1959] "causally relevant factor" account. A refinement of the NESS/INUS test that integrates suggestions from the causal model approach of Halpern and Pearl [2005] has been recently proposed by Baumgartner [2013].

formed by agents). The analysis is only carried out at the semantic level and agency is not represented explicitly.

3. Belnap and Perloff [1993] use the notion of minimal sufficient condition (but not that of NESS condition) to analyze a notion of group agency called "strict joint agency" in standard STIT semantics.

4. Bulling and Dastani [2013] and Lorini et al. [2014] introduce notions of responsibility, respectively in CL and STIT, defined in terms of the power of an agent to prevent a certain state of affairs. An agent has the power to prevent φ if *there is* an action available to him that guarantees the truth of $\neg\varphi$. This notion is intuitively weaker than that of *but-for* condition: roughly, what an agent does is a *but-for* condition of φ if *all* alternative actions available to him guarantee the truth of $\neg\varphi$ (more on this in Section 3.1.3). In both Bulling and Dastani [2013] and Lorini et al. [2014] action labels that can be interpreted as action types are present in the semantics but not in the object language.

5. In order to analyze "unwitting" group agency, Sergot [2008] studies so-called counteraction conditions of different strength that approximate the *but-for* test. Sergot's proposal is in the tradition of "bring it about" logics [esp. Pörn, 1977] rather than STIT. Still, the semantical framework he advances is a form of labeled transition system, where properties of transitions that can be interpreted as action types can be expressed in the object language.

6. Braham and van Hees [2011, 2012, 2018] provide a game-theoretic formulation of the NESS test to analyze moral responsibility and responsibility voids. Our formalization of the NESS test in Section 3.1.3 can be viewed as a recasting of Braham and van Hees's formulation in STIT.

7. By using the machinery of *individual* strategic STIT from Belnap et al. [2001, Chapter 13], Müller [2005] provides a semantics for formulas like $[i\,istit]\varphi$, read "agent i is seeing to it that φ." Here φ is the *default result* of the *concrete strategy* that agent i is performing. Suggestions to interpret such formulas in a STIT semantics supplied with action types are advanced by Troquard and Vieu [2006], who, nevertheless, leave the formalization of the notion of default result to future work. Conceptually, the main difference between Müller's proposal and our own is that Müller assumes a primitive notion of "default strategy" and takes default results to be the results of this strategy. On the other hand, by taking a cue from the literature on causal models, we assume a primitive notion of external interference and take expected results to be the results that an action would have if performed without interference.

With respect to these works, our main contributions are to supplement STIT with notions of actual causality that are well-established in the legal and philosophical literature and to make them interact with a notion of potential causality

in order to analyze causal responsibility. As we will see in Section 3.4, both potential and actual causality are crucial to capture important aspects of individual and group responsibility. Finally, even though STIT has been extended with action types in number of earlier works [see page 32], the introduction of a relation between action types to model causal influences between them is new.

3.2 The action logic with opposing ALO_n

In this section, we introduce a STIT logic with action types and a relation of opposing between them. We start, in Section 3.2.1, by introducing the ingredients that are needed to give the semantics for the logic. The formal language, the models, and the notion of truth are defined in Section 3.2.2, where a complete axiomatization is also presented (the proof of completeness is in Appendix A.1).

3.2.1 ALO_n frames

Our semantical framework consists of three main components: a *discrete branching time structure* (called DBT structure) modeling the flow of time; an *action type function* labeling moment-history pairs with the action types performed by the group of all agents; finally, an *opposing function* representing the relation of opposing between action types.

3.2.1. Definition (DBT structure). A DBT structure is a tuple $\langle Mom, <\rangle$ such that $Mom \neq \varnothing$ is a set of moments and $<$ is the temporal precedence relation. As usual, $\leq \subseteq Mom \times Mom$ is defined as: for any $m, m' \in Mom$: $m \leq m'$ if and only if $m < m'$ or $m = m'$. The relation $<$ is a discrete tree-like ordering of Mom: it satisfies, for all $m, m_1, m_2, m_3 \in Mom$,

1. *Irreflexivity*: $m \not< m$.
2. *Transitivity*: if $m_1 < m_2$ and $m_2 < m_3$, then $m_1 < m_3$.
3. *Past-linearity*: if $m_1 \leq m_3$ and $m_2 \leq m_3$, then either $m_1 \leq m_2$ or $m_2 \leq m_1$.
4. *Discreteness*: if $m_1 < m_2$, then there is an m_3 such that $m_1 < m_3 \leq m_2$ and there is no m_4 such that $m_1 < m_4 < m_3$.
5. *No endpoints*: there is an $m' \in Mom$ such that $m < m'$.

The standard notions used to reason about DBT structures are summarized in Table 3.1. We extensively discussed the notions in groups (I) and (III) in Chapter 2.1. To summarize the key points and add the notions in group (II), let $\mathcal{T} = \langle Mom, <\rangle$ be a DBT structure. Recall that each $h \in Hist^{\mathcal{T}}$ represents a complete course of events. If $m \in Mom$, then each $h \in H_m^{\mathcal{T}}$ represents a complete

(I) Histories

- A *history* is a maximal set of linearly ordered moments from Mom.
- $Hist^{\mathcal{T}}$ is the set of histories in $\mathcal{T}$.
- History h *passes through moment* m when $m \in h$.
- $H_m^{\mathcal{T}} = \{h \in Hist^{\mathcal{T}} \mid m \in h\}$ is the set of histories passing through m.
- $h, h' \in Hist^{\mathcal{T}}$ are *undivided* at m iff $m \in h \cap h'$ and there is $m' > m$ s.t. $m' \in h \cap h'$.

(II) Immediate successors

- $succ(m) = \{m' \in Mom \mid m < m'$ and, for no $m'' \in Mom, m < m'' < m'\}$ is the set of *immediate successors* of m.
- If $h \in H_m^{\mathcal{T}}$, the *immediate successor of* m *on* h, denoted with $succ_h(m)$, is the unique element of $h \cap succ(m)$.

(III) Indices

- An *index* is a pair m/h such that $m \in Mom$ and $h \in H_m^{\mathcal{T}}$.
- $Ind^{\mathcal{T}}$ is the set of indices in $\mathcal{T}$.

Table 3.1: Key notions related to a DBT structure $\mathcal{T}$

course of events that can still be realized at m and the moment-history pair m/h represents the complete state of the world at moment m on history h. Since time is discrete with no endpoint, for each $m \in Mom$, the set $succ(m)$ of immediate successors of m is non-empty. In addition, if $h \in H_m^{\mathcal{T}}$, then $h \cap succ(m)$ is a singleton because histories are linearly ordered sets of moments. The unique element of this set is the immediate successor of moment m on history h and is denoted with $succ_h(m)$. As before, we will omit the superscript $\mathcal{T}$ and simply write $Hist$, H_m, and Ind when the DBT structure is clear from context.

Turning to agency, let us start by fixing sets of agents and action types:

- Let $Ag = \{1, \ldots, n\}$ be the set of n (names of) *agents* for some number $n \in \mathbb{N}$. We will use $i, j, k, i, i'', \ldots$ for elements of Ag.
- Let Atm be a non-empty finite set of (names of) *action types*. We will use $a, b, c, a', a'', \ldots$ for elements of Atm.

We think of agents as being endowed with a repertoire of action types of which they can be authors. We associate each action type with its possible authors and fix a set $Acts$ of *individual action types*, defined as follows:

$$Acts \subseteq Atm \times Ag$$

We write a_i rather than (a, i) when $(a, i) \in Acts$. Intuitively, a_i is the action type that is instantiated whenever agent i performs an action of type a. For instance,

if a is the action type *writing* and $1, 2 \in Ag$ are, respectively, Tom and Aybüke, then a_1 is the action type *Tom's writing* and a_2 is the action type *Aybüke's writing*. For $i \in Ag$, let $Acts_i$ be the set of action types authored by agent i:

$$Acts_i = \{a_i \in Acts \mid a \in Atm\}.$$

A *complete group action* is a function $\alpha : Ag \to Acts$ such that, for all $i \in Ag$, $\alpha(i) \in Acts_i$. So, a complete group action is any combination of individual actions, one for each agent (in game-theoretic terms, it is an *action profile*). Let *Ag-Acts* be the set of all complete group actions (we use $\alpha, \beta, \gamma, \ldots$ for elements of *Ag-Acts*). Intuitively, $\alpha \in$ *Ag-Acts* is the action type that is instantiated whenever, for all $i \in Ag$, agent i performs action $\alpha(i)$. As usual, when $\alpha \in$ *Ag-Acts* and $I \subseteq Ag$, we will write α_I for the restriction of α to the set I, α_{-I} for $\alpha_{Ag \setminus I}$, and $\alpha(I)$ for the image of I under α. For any $\alpha \in$ *Ag-Acts* and $I \subseteq Ag$, α_I is a *group action*. Let *G-Acts* be the set of all group actions. We define a relation $\sqsubseteq \subseteq$ *G-Acts* $\times$ *G-Acts* by setting: for all $\alpha_I, \beta_J \in$ *G-Acts*,

$$\alpha_I \sqsubseteq \beta_J \text{ iff } \alpha(I) \subseteq \beta(J).$$

$\alpha_I \sqsubseteq \beta_J$ means that α_I is a *sub-action* of β_J, or that α_I is included in β_J.

We take the action types in $Atm, Acts$, *Ag-Acts*, and *G-Acts* to represent *one-step actions*. So, in the spirit of PDL [Harel et al., 2000] and CL [Pauly, 2002], performing an action at a moment transitions to a set of *next* moments representing the different possible outcomes of the action.[16] To make this explicit, we could define a *transition* in a DBT structure $\mathcal{T}$ to be any pair of moments (m, m') such that $m' \in succ(m)$, where m is the initial-moment of the transition and m' is its end-moment. As in game models for CL, we could then label transitions with the complete group actions that, intuitively, bring them about.[17] But in order to avoid the introduction of further notation, here we will label indices instead: every index m/h will be labeled with the complete group action that, intuitively, brings about the transition from m to its successor on h (i.e., the moment $succ_h(m)$).[18] If index m/h is labeled with $\alpha \in$ *Ag-Acts*, then $\alpha(i)$ represents *the action type that agent i instantiates at m/h* and, similarly, α_I the action type that group I instantiates at m/h. Hence, every agent i instantiates one, and only one, type of action at every index m/h. We take this to stand for the action type that the modeler has selected in order to describe what agent i is doing at m/h. As discussed in Section 3.1.1, which type is selected and how specific the selected type is depend on the purposes of the modeler.

The final component of the semantics is an *opposing function* that assigns to every individual or group action the set of individual or group actions that oppose

[16] We think of the assumption that the temporal ordering is discrete as a by-product of this view, rather than as an assumption about the structure of time in itself.

[17] This is what we do in Baltag et al. [Forthcoming].

[18] Unlike action profiles in game models, complete group actions in our frames will not necessarily determine a *unique* transition.

it.[19] An individual or group action opposes another individual or group action when the performance of the former is generally assumed to interfere with the performance of the latter. Since opposing is an intrinsic relation between action types [see Section 3.1.2], the opposing function is not moment-relative. Unless differently specified, we will use "action" as a shortcut for "individual or group action" in what follows.

3.2.2. Definition (ALO_n frame). An ALO_n frame is a tuple $\langle \mathcal{T}, \mathbf{act}, \mathbf{opp} \rangle$, where $\mathcal{T}$ is a DBT structure, $\mathbf{act} : \mathit{Ind} \to \mathit{Ag\text{-}Acts}$ assigns to every index the complete group action that is performed at it, and $\mathbf{opp} : \mathit{G\text{-}Acts} \to 2^{\mathit{G\text{-}Acts}}$ assigns to every action the set of actions opposing it. For any $m \in \mathit{Mom}$ and $i \in \mathit{Ag}$, let

$$\mathit{Acts}_i^m = \bigcup_{h \in H_m} \mathbf{act}(m/h)(i)$$

be the set of *actions available to agent i at m* and

$$\mathit{Acts}^m = \bigcup_{i \in \mathit{Ag}} \mathit{Acts}_i^m$$

the set of *individual actions executable at m*. The functions **act** and **opp** are required to satisfy the following conditions: for all $m \in \mathit{Mom}$, $h, h' \in \mathit{Hist}$, and $\alpha_I, \beta_J, \gamma_K \in \mathit{G\text{-}Acts}$,

1. *No Choice Between Undivided Histories*: if h and h' are undivided at m, then $\mathbf{act}(m/h) = \mathbf{act}(m/h')$.
2. *Independence of Agents*: for all $\alpha \in \mathit{Ag\text{-}Acts}$, if $\alpha(j) \in \mathit{Acts}^m$ for all $j \in \mathit{Ag}$, then there is $h \in H_m$ such that $\mathbf{act}(m/h) = \alpha$.
3. *Irreflexivity of Opposing*: $\alpha_I \notin \mathbf{opp}(\alpha_I)$.
4. *Monotonicity of Opposing*: if $\alpha_I \in \mathbf{opp}(\beta_J)$ and $\beta_J \sqsubseteq \gamma_K$, $\alpha_I \in \mathbf{opp}(\gamma_K)$.

It is not difficult to see that the set Acts_i^m of actions available to agent i at moment m induces a partition on H_m: for every $h \in H_m$, the set

$$\mathit{Acts}_i^m(h) = \{h' \in H_m \mid \mathbf{act}(m/h)(i) = \mathbf{act}(m/h')(i)\}$$

is the cell in the partition containing h. The set $\mathit{Acts}_i^m(h)$ is the action token performed by i at m/h familiar in STIT semantics [see Chapter 2.1.1] that has been tagged with its assigned type. Note that every such action token is assigned a unique type and different tokens are assigned different types.[20]

[19] In defining this function, we will use the set *G-Acts* instead of the set $\mathit{Acts} \cup \mathit{G\text{-}Acts}$. The reason is that individual actions can be represented by means of singleton-group actions.

[20] This is a common idea and can be found in, e.g., Horty and Pacuit [2017]. It is also at the basis of the proof, presented by Broersen et al. [2006b], that CL [Pauly, 2002] can be embedded in STIT. See Chapter 2.3.1 for more details.

Conditions 1 and 2 from Definition 3.2.2 are standard requirements in STIT semantics [cf. Chapter 2.1]: The condition of no choice between undivided histories ensures that no individual action executable at a moment can separate histories that are undivided at that moment. The condition of independence of agents ensures that every combination of individual actions executable at a moment (one for each agent) can itself be executed at that moment.

Turning to the conditions on the function **opp** (conditions 3 and 4 from Definition 3.2.2), irreflexivity and monotonicity of opposing ensure that no action opposes itself and that opposing a sub-action suffices to oppose the action as a whole, in accordance with our basic view on the opposing relation [see Section 3.1.2]. Note that these two requirements entail the following condition, according to which no action opposes any of its sub-actions:

5. For any $\alpha_I, \beta_J \in \textit{G-Acts}$, if $\alpha_I \sqsubseteq \beta_J$, then $\beta_J \notin \mathbf{opp}(\alpha_I)$.

To see this, suppose, toward contradiction, that $\alpha_I \sqsubseteq \beta_J$ and $\beta_J \in \mathbf{opp}(\alpha_I)$. Then, by monotonicity of opposing, $\beta_J \in \mathbf{opp}(\beta_J)$, against irreflexivity of opposing. Hence, $\beta_J \notin \mathbf{opp}(\alpha_I)$. Condition 5 reflects the idea that, by opposing its parts, an action would oppose itself. Importantly, this condition does not exclude the possibility that a sub-action of an action α_I opposes another sub-action of α_I. This is a desirable feature: *Alice's shooting Dan and Beth's hitting Alice* is an example (among many others) of a group action consisting of two actions one of which opposes the other.

3.2.2 Syntax, semantics, and axiomatization

We now introduce the language $\mathcal{L}_{\mathsf{ALO}_n}$ of the logic ALO_n. We start by fixing, besides the set $Ag = \{1, \ldots, n\}$ of (names of) agents and the set Atm of (names of) action types [see p. 48], a non-empty countable set $Prop$ of propositional variables.

3.2.3. Definition (Syntax of $\mathcal{L}_{\mathsf{ALO}_n}$). Let Ag, Atm, and $Prop$ be defined as above. The set of formulas of $\mathcal{L}_{\mathsf{ALO}_n}$, also denoted with $\mathcal{L}_{\mathsf{ALO}_n}$, is generated by the following grammar:

$$\varphi := \; p \,|\, do(a_i) \,|\, \alpha_I \triangleright \beta_J \,|\, \neg\varphi \,|\, (\varphi \wedge \varphi) \,|\, \Box\varphi \,|\, \mathsf{X}\varphi$$

where $p \in Prop$, $a_i \in Acts$, and $\alpha_I, \beta_J \in \textit{G-Acts}$. The abbreviations for the Boolean connectives $\vee$, $\rightarrow$, $\leftrightarrow$, for $\bot$ and $\top$ are standard. As usual, we use $\Diamond\varphi$ and $\hat{\mathsf{X}}\varphi$ as abbreviations for $\neg\Box\neg\varphi$ and $\neg\mathsf{X}\neg\varphi$ respectively. In addition, for any $I \subseteq Ag$ and $\alpha \in \textit{Ag-Acts}$, we introduce the following abbreviations:[21]

$$do(\alpha_I) := \bigwedge_{a_i \in \alpha(I)} do(a_i) \qquad\qquad do(\alpha) := \bigwedge_{a_i \in \alpha(Ag)} do(a_i)$$

[21]Since both Ag and Atm are finite, both $\bigwedge_{a_i \in \alpha(I)} do(a_i)$ and $\bigwedge_{a_i \in \alpha(Ag)} do(a_i)$ are finite conjunctions, therefore, guaranteed to be in the language.

Finally, we will follow the usual rules for the elimination of parentheses.

3.2.4. Remark. Following a common practice in game theory and CL, in Section 3.2.1 we have introduced (complete) group actions as *functions* assigning to every agent from a given group one of the individual actions in her repertoire. A perhaps less intuitive – but technically more rigorous – way to go is to define *Ag-Acts* as the *set* of individual actions that satisfies the following condition: for every $\alpha \subseteq Acts$, $\alpha \in Ag\text{-}Acts$ iff, for all $i \in Ag$, $Acts_i \cap \alpha$ is a singleton. The following notation could then be introduced: 1) for any $\alpha \in Ag\text{-}Acts$ and $i \in Ag$, $\alpha(i)$ is the unique element of $Acts_i \cap \alpha$; 2) for every $\alpha \in Ag\text{-}Acts$ and $I \subseteq Ag$, $\alpha_I = \{a_i \in \alpha \mid i \in I\}$ (and similarly for $\alpha(I)$); finally, 3) the sub-action relation $\sqsubseteq$ is just set inclusion. This is what we have in mind when we use (complete) group actions and the related notation in the syntax of ALO_n.

The language $\mathcal{L}_{\mathsf{ALO}_n}$ includes the temporal operator X for what happens next, the operator $\Box$ of historical necessity, and the action formulas built from *do* and $\triangleright$. As usual, $\mathsf{X}\varphi$ means "φ is true next" and is true at an index m/h whenever φ is true at the immediate successor of m on h; $\Box\varphi$ means "φ is settled true" or "φ is historically necessary" and is true at an index m/h whenever φ is true at m on all histories passing through it. We read a formula like $do(a_i)$ as "agent i is doing an action of type a" and take it to be true at an index m/h whenever agent i performs an action of type a_i at that index. A formula like $\alpha_I \triangleright \beta_J$ means "action α_I opposes action β_J," and is taken to be true at an index m/h whenever α_I is among the actions that oppose β_J.

The reading of the action predicate *do* is particularly important. We understand it as expressing *continuous actions* (*being in the process of doing something*) that can fail or be interrupted before completion. Accordingly, a formula like $do(a_i)$ means that agent i is carrying out an action of type a, without this implying anything about his success or failure in doing so. The important point is that it makes perfect sense to say that an agent is doing something even if, in the end, she fails because of some external opposition.[22]

Modalities in the spirit of PDL and XSTIT can be expressed in $\mathcal{L}_{\mathsf{ALO}_n}$ as follows:[23]

3.2.5. Definition (PDL and XSTIT modalities). Where $\alpha \in Ag\text{-}Acts$, $I \subseteq Ag$, and $\varphi \in \mathcal{L}_{\mathsf{ALO}_n}$,

$$[\alpha_I]\varphi := \Box(do(\alpha_I) \to \mathsf{X}\varphi) \qquad [I\,xstit]\varphi := \bigvee_{\alpha \in Ag\text{-}Acts}(do(\alpha_I) \wedge [\alpha_I]\varphi)$$
$$\langle\alpha_I\rangle\,\varphi := \Diamond(do(\alpha_I) \wedge \mathsf{X}\varphi) \qquad [I\,dxstit]\varphi := [I\,xstit]\varphi \wedge \neg\Box\mathsf{X}\varphi$$

[22] Recall the examples from Section 3.1.2: my flatmate is coming upstairs, but she might fail because I am going downstairs at the same time; my sister is watering the plants, but she might fail because her partner trips over the garden hose; and so on.

[23] Similar definitions can be found in, e.g., Broersen [2014b], Herzig and Lorini [2010], and Segerberg [2002].

Accordingly, $[\alpha_I]\varphi$ means that, no matter how the future unfolds, doing α_I ensures that φ will be the case next and, dually, $\langle\alpha_I\rangle\,\varphi$ means that the future might unfold in such a way that α_I is done and φ is true next. Thus, $\langle\alpha_I\rangle\top$ expresses that group I can perform an action of type α_I. Turning to the XSTIT operators, $[I\,xstit]\varphi$ says that group I is doing an action that ensures that φ is true in the next moment. The deliberative XSTIT operator $[I\,dxstit]$ is defined as usual by adding to the STIT formula $[I\,xstit]\varphi$ the requirement that φ is not already settled [see Section 2.1.2].

The notions of ALO_n model and truth of formulas from $\mathcal{L}_{\mathsf{ALO}_n}$ are defined as follows.

3.2.6. Definition (ALO_n model). Let *Prop* be defined as above. An ALO_n model is a tuple $\langle\mathcal{F},\pi\rangle$, where $\mathcal{F}$ is an ALO_n frame and $\pi : \mathit{Prop} \to 2^{\mathit{Ind}}$ is a valuation function.

3.2.7. Definition (Semantics for $\mathcal{L}_{\mathsf{ALO}_n}$). Given an ALO_n model $\mathcal{M}$, truth of a formula $\varphi \in \mathcal{L}_{\mathsf{ALO}_n}$ at an index m/h in $\mathcal{M}$, denoted $\mathcal{M}, m/h \models \varphi$, is defined recursively. Truth of atomic propositions and the Boolean connectives is defined as usual. The remaining clauses are as follows:

$$\begin{array}{lll}
\mathcal{M}, m/h \models \mathit{do}(a_i) & \text{iff} & \mathbf{act}(m/h)(i) = a_i \\
\mathcal{M}, m/h \models \alpha_I \triangleright \beta_J & \text{iff} & \alpha_I \in \mathbf{opp}(\beta_J) \\
\mathcal{M}, m/h \models \mathsf{X}\varphi & \text{iff} & \mathcal{M}, \mathit{succ}_h(m)/h \models \varphi \\
\mathcal{M}, m/h \models \Box\varphi & \text{iff} & \text{for all } h' \in H_m,\ \mathcal{M}, m/h' \models \varphi
\end{array}$$

Notice that, as an immediate consequence of Definition 3.2.7, formulas like $[I\,xstit]\varphi$ have the following truth condition, in line with the standard semantics for group STIT modalities:

$$\begin{array}{lll}
\mathcal{M}, m/h \models [I\,xstit]\varphi & \text{iff} & \mathcal{M}, m/h \models \bigvee_{\alpha\in \mathit{Ag\text{-}Acts}}(\mathit{do}(\alpha_I) \wedge \Box(\mathit{do}(\alpha_I) \to \mathsf{X}\varphi)) \\
& \text{iff} & \text{there is } \alpha \in \mathit{Ag\text{-}Acts} \text{ s.t., for all } a_i \in \alpha(I),\ \mathbf{act}(m/h)(i) = a_i \\
& & \quad \text{and, for all } h' \in H_m \text{ s.t., for all } a_i \in \alpha(I),\ \mathbf{act}(m/h')(i) = a_i, \\
& & \quad \mathcal{M}, m/h' \models \mathsf{X}\varphi \\
& \text{iff} & \text{for all } h' \in \bigcap_{i\in I} \mathit{Acts}_i^m(h),\ \mathcal{M}, m/h' \models \mathsf{X}\varphi
\end{array}$$

The proof of the following theorem can be found in Appendix A.1.

3.2.8. Theorem. *The axiom system* ALO_n *defined by the axioms and rules in Table 3.2 is sound and complete with respect to the class of all* ALO_n *frames.*

The axiom system ALO_n extends a standard axiomatization for the modalities $\Box$ and X with bridge principles connecting these modalities to the action description formulas using *do* and $\triangleright$. The axioms for *do* are a reformulation, in $\mathcal{L}_{\mathsf{ALO}_n}$, of the main axioms of the Dynamic Logic of Agency ($\mathcal{DLA}$) proposed by Herzig

(CPL)	Classical propositional tautologies	(MP)	From φ and $\varphi \to \psi$, infer ψ
($\mathsf{S5}_\Box$)	The axiom schemas of S5 for $\Box$	($\mathsf{RN}_\Box$)	From φ, infer $\Box\varphi$
(KD_X)	The axiom schemas of KD for X	(RN_X)	From φ, infer $\mathsf{X}\varphi$
(F_X)	$\hat{\mathsf{X}}\varphi \to \mathsf{X}\varphi$		
	(I) Axioms for *do*:		
(UH_{do})	$(do(\alpha) \wedge \mathsf{X}\Diamond\varphi) \to \Diamond(do(\alpha) \wedge \mathsf{X}\varphi)$	(Act)	$\bigvee_{a_i \in Acts_i} do(a_i)$
(IA_{do})	$(\Diamond do(a_1) \wedge \cdots \wedge \Diamond do(a_n)) \to \Diamond do(\alpha)$	(Sin)	$do(a_i) \to \neg do(b_i)$
	for $\alpha(1) = a_1, \ldots, \alpha(n) = a_n$		*for* $a_i \neq b_i$
	(II) Axioms for $\triangleright$:		
($\mathsf{Irr}_\triangleright$)	$\neg(\alpha_I \triangleright \alpha_I)$	($\mathsf{Sett}_\triangleright$)	$\alpha_I \triangleright \beta_J \to \Box(\alpha_I \triangleright \beta_J)$
($\mathsf{Mon}_\triangleright$)	$\alpha_I \triangleright \beta_J \to \alpha_I \triangleright \gamma_K$	($\mathsf{Fix}_\triangleright$)	$\alpha_I \triangleright \beta_J \leftrightarrow \mathsf{X}(\alpha_I \triangleright \beta_J)$
	provided that $\beta_J \sqsubseteq \gamma_K$		

Table 3.2: The axiom system ALO_n

and Lorini [2010].[24] Axiom UH_{do} expresses the constraint of no choice between undivided histories, according to which what the group of all agents presently does cannot exclude that what might happen next will indeed happen next.[25] Axiom IA_{do} expresses the constraint of independence of agents: if the individual actions $a_1, \ldots, a_n$ can be performed separately, then these actions can also be performed jointly. Finally, Axioms Act (for "Active") and Sin (for "Single") say that every agent performs one, and only one, action at every index. The remaining axioms characterize the opposing relation. While $\mathsf{Irr}_\triangleright$ and $\mathsf{Mon}_\triangleright$ correspond to the properties of irreflexivity and monotonicity of opposing (i.e., conditions 3 and 4 from Definition 3.2.2), $\mathsf{Sett}_\triangleright$ and $\mathsf{Fix}_\triangleright$ reflect the fact that the opposing relation is modeled by a global function: if action α_I opposes another action β_J, then it is settled that α_I opposes β_J; what is more, α_I will always oppose and has always opposed β_J.

3.3 Causal agency and responsibility in ALO_n

As discussed in Section 3.1, we typically determine who, among a set of relevant agents performing some relevant types of action, is causally responsible for a

[24] It can be proved that there is a double embedding between the fragment of ALO_n without the opposing operator and $\mathcal{DLA}$.

[25] Axiom UH_{do} is analogous to the axiom UH of the axiom system STIT_n^{Ag} that we discussed in Section 2.2.3.

state of affairs φ by first identifying the actions that potentially caused φ and then selecting, among them, those that actually contributed to it. According to an intuitive view, we identify the potential causes of φ by considering the actions that were *expected to result* in φ. We then single out the actual causes of φ by checking which of its potential causes passes a given test for actual causation. Here we will rely on the two tests that are the most widely accepted in the legal literature, namely the *but-for test* and the *NESS tests*. The logic ALO_n is a fairly simple and, yet, powerful refinement of STIT logic that can be used to formalize this commonsense view. In this section, we begin by providing a representation of the key notions of expected result, *but-for* dependence and NESS dependence, and by defining new STIT operators modeling causal responsibility (we will use "responsibility" as a shortcut for "causal responsibility" from now on). We will apply these notions to the analysis of concrete example in the next Section 3.4.

3.3.1 Expected-result conditionals

In Section 3.1.2 we suggested that a common way to test whether φ is an expected result of an action is to consider scenarios in which nothing opposes the execution of that action and see whether φ is produced. We introduced the relation of opposing in our framework precisely to be able to capture this view. Our starting point is the following notion of *unopposed execution* of an action $\alpha_I \in G\text{-}Acts$:

3.3.1. Definition (Doing unopposed). Where $\alpha_I \in Ag\text{-}Acts$,

$$\underline{do}(\alpha_I) := do(\alpha_I) \wedge \bigwedge_{\beta_J \in G\text{-}Acts} (do(\beta_J) \rightarrow \neg(\beta_J \triangleright \alpha_I))$$

According to Definition 3.3.1, action α_I is done unopposed just in case α_I is performed and no other performed action opposes it. In the semantics, this is the same as saying that no sub-action of the complete group action that is performed opposes α_I. To see this, where $\alpha_I \in G\text{-}Acts$ and $\gamma \in Ag\text{-}Acts$, let:

$$\mathbf{unopp}(\alpha_I, \gamma) \text{ iff } \alpha_I \sqsubseteq \gamma \text{ and there is no } J \subseteq Ag \text{ s.t. } \gamma_J \in \mathbf{opp}(\alpha_I) \qquad (1)$$

$\mathbf{unopp}(\alpha_I, \gamma)$ (read "α_I *is unopposed in* γ") says that none of the sub-actions of γ opposes α_I. It is not difficult to check that $\underline{do}(\alpha_I)$ has the following interpretation in ALO_n models:

$$\mathcal{M}, m/h \models \underline{do}(\alpha_I) \text{ iff } \mathbf{unopp}(\alpha_I, \mathbf{act}(m/h)) \qquad (2)$$

We can now define what it means that φ *is an expected results of an action* α_I in a simple way:

3.3.2. Definition (Expected result). Where $\alpha_I \in G\text{-}Acts$ and $\varphi \in \mathcal{L}_{\mathsf{ALO}_n}$,

$$do(\alpha_I) \boxplus\!\!\rightarrow \varphi := \Box(\underline{do}(\alpha_I) \rightarrow \mathsf{X}\varphi)$$

Definition 3.3.2 expresses that we expect α_I to result in φ when φ would be the case if α_I were done unopposed. It is immediate to see that $do(\alpha_I) \boxplus\!\!\rightarrow \varphi$ has the following interpretation in ALO_n models:

$$\mathcal{M}, m/h \models do(\alpha_I) \boxplus\!\!\rightarrow \varphi \text{ iff, for all } h' \in H_m \text{ s.t. } \mathbf{unopp}(\alpha_I, \mathbf{act}(m/h')), \mathcal{M}, m/h' \models \mathsf{X}\varphi \tag{3}$$

The semantic clause 3 makes evident the counterfactual flavor of the notion of expected result: $do(\alpha_I) \boxplus\!\!\rightarrow \varphi$ is true at index m/h just in case, in all hypothetical scenarios in which α_I is done unopposed at m, φ is produced at the successor of m.[26] In this sense, complete group actions that can be performed at m and in which α_I is unopposed can be seen as flags indicating the hypothetical scenarios that we need to consider if we want to carry out the expected-result test.

Let us illustrate the notions introduced so far with an example.

3.3.3. Example. Alice is closing her shop. Beth does not want to leave and grabs her. At the same time, Carl decides to try to enter the shop instead of going home. Since Beth is holding Alice, the door remains open and Carl enters.

3.3.4. Remark. In what follows we will adopt the following notation. For all $\alpha_I, \beta_J \in G\text{-}Acts$ such that $I \cap J \neq \varnothing$, $\alpha_I\beta_J$ is the group action induced by α_I and β_J: that is, for all $i \in I \cup J$, $\alpha_I\beta_J(i) = \alpha(i)$ if $i \in I$ and $\alpha_I\beta_J(i) = \beta_J(i)$ if $i \in J$. In addition, when $\alpha_{\{i\}}(i) = a_i$, we will often write $\mathbf{opp}(a_i)$ instead of $\mathbf{opp}(\alpha_{\{i\}})$.

Example 3.3.3 can be represented as the ALO_n model depicted in Figure 3.2(A), where Alice is agent 1 and can perform either action cs_1 (closing the shop) or action so_1 (leaving the shop open), Beth is agent 2 and can perform either action ga_2 (grabbing Alice) or action ls_2 (leaving the shop), and Carl is agent 3 and can perform either action es_3 (entering the shop) or action gh_3 (going home). p stands for the proposition "Carl is in the shop." For the sake of readability, the actions available to the three agents at m_1 are represented as tags of the choice-cells that they induce (rather than as index-labels): the actions available to Alice tag the columns, those available to Carl the rows, and those available to Beth the areas within and outside the circle. The actual scenario corresponds to index m_1/h_1 where Alice closes the shop, Beth grabs her, and Carl tries to enter and succeeds.

At the actual index m_1/h_1, neither Carl's nor Alice's action is done unopposed: Carl's action of trying to enter the shop is opposed by Alice's action of closing it; in turn, Alice's action of closing the shop is opposed by Beth's action of grabbing her. On the other hand, since no action opposes Beth's action of grabbing Alice,

[26] Let us emphasize that the "expected-result conditional" $\boxplus\!\!\rightarrow$ is defined as a *strict* rather than as *counterfactual* conditional. So, although the notion of expected result has a counterfactual flavor, $\boxplus\!\!\rightarrow$ does not have the properties of counterfactual conditionals in the tradition of Stalnaker [1968] and Lewis [1973a]. We will discuss the problem of merging STIT with a logic for counterfactuals in Chapter 4.

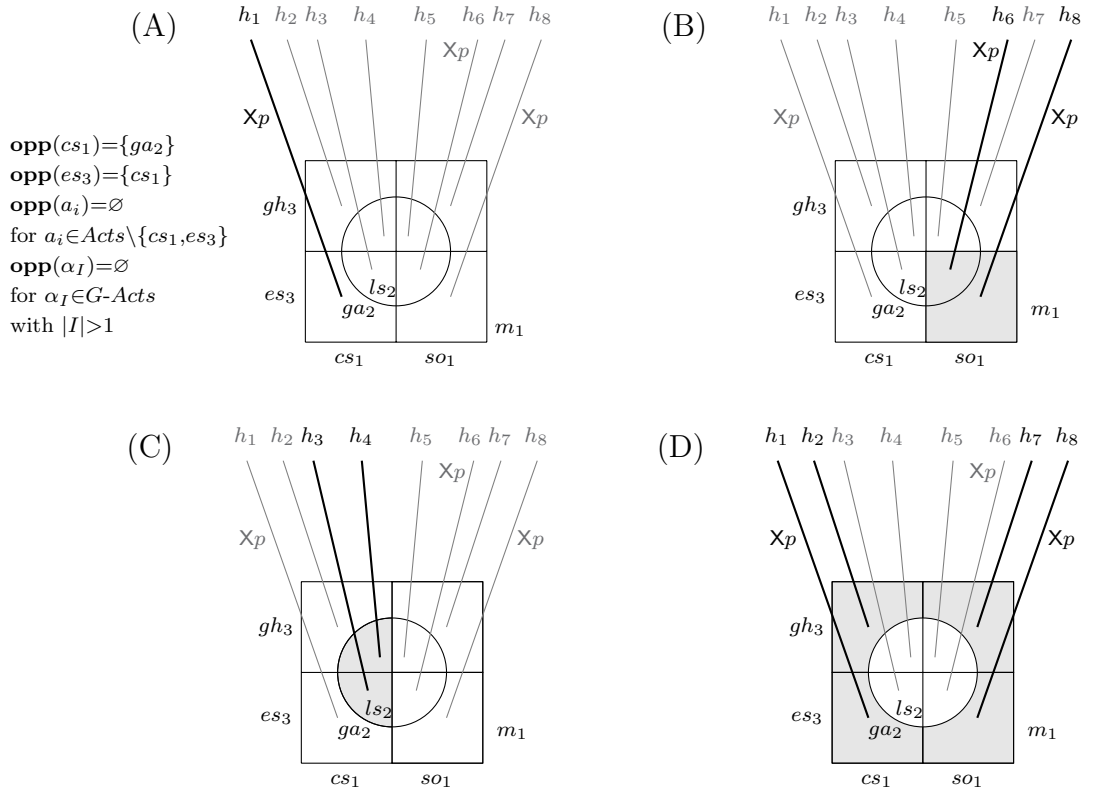


Figure 3.2: ALO_n models representing Example 3.3.3

the latter action is done unopposed. The gray shaded area in Figure 3.2(B) identifies the complete group actions in which es_3 is unopposed, namely all actions consisting of the types *Alice's leaving the shop open* and *Carl's entering* (i.e., $so_1 ga_2 es_3$ and $so_1 ls_2 es_3$). Since $\mathsf{X}p$ is true on all histories on which these actions are performed, $do(es_3) \boxplus\!\!\rightarrow p$ is true at m_1/h_1: Carl's action of entering the shop is expected to result in a state in which Carl is in the shop. The gray shaded area in Figure 3.2(C) identifies the complete group actions in which cs_1 is unopposed (the actions consisting of the types *Alice's closing the shop* and *Beth's leaving*, i.e., $cs_1 ls_2 es_3$ and $cs_1 ls_2 gh_3$) and the gray shaded area in Figure 3.2(D) the complete group actions in which ga_2 is unopposed (all complete group actions including ga_2). By inspecting the two Figures 3.2(C) and 3.2(D) it is easy to see that $do(cs_1) \boxplus\!\!\rightarrow \neg p$ is true at m_1/h_1, while neither $do(ga_2) \boxplus\!\!\rightarrow p$ nor $do(ga_2) \boxplus\!\!\rightarrow \neg p$ are true at m_1/h_1: while Alice's action of closing the shop is expected to result in a state in which Carl is not in the shop, Beth's action of grabbing Alice is not expected to determine whether Carl will be in the shop or not.

3.3.2 Tests for actual causation

In analyzing Example 3.3.3, we have seen that Carl's action was expected to result in Carl's being inside the shop, but which actions did *actually contribute* to this result? According to the *but-for* test for actual causation, any action that was such that, *but for* that action, Carl would not have been in the shop; that is, any action that, *in that specific situation*, was necessary for Carl's being inside. If we read the phrase "in that specific situation" as "*given what the other agents were doing*," we can formalize what it means that an action α_I *is a but-for cause of* φ as follows:

3.3.5. Definition (*But-for* cause). Where $\alpha_I \in \textit{G-Acts}$ and $\varphi \in \mathcal{L}_{\mathsf{ALO}_n}$,

$$\textit{but}(\alpha_I, \varphi) := \mathsf{X}\varphi \wedge \bigvee_{\substack{\gamma \in \textit{Ag-Acts}: \\ \alpha_I \sqsubseteq \gamma}} (\textit{do}(\gamma) \wedge \bigwedge_{\substack{\beta \in \textit{Ag-Acts}: \\ \beta_I \neq \alpha_I}} [\gamma_{-I}\beta_I]\neg\varphi)$$

According to Definition 3.3.5, α_I is a *but-for* cause of φ just in case φ happens next and the complete group action that is performed includes α_I and is such that no complete group action obtained from it by replacing α_I with any other action for group I results in φ. Intuitively, this means that, other things being equal, φ would not have been realized had α_I not been performed.

Let us go back to Example 3.3.3 and Figure 3.2 to illustrate how the definition works. It is immediate to see that $\mathsf{X}p \wedge \textit{do}(cs_1ga_2es_3)$ is true at the actual index m_1/h_1. The question is what are the *but-for* causes of p (if any). Alice is not one of them: Given that Carl tries to enter the shop, he would end up inside even if Alice decided to keep it open. This is captured by the fact that $[so_1ga_2es_3]p$ is true at m_1/h_1, which makes $\textit{but}(cs_1, p)$ false at this index. On the other hand, given that Alice is closing the shop, Carl would not enter if Beth decided to leave (recall that $\neg p$ is an expected result of cs_1). This is reflected by the fact that $[cs_1ls_2es_3]\neg p$ is true at m_1/h_1, which makes $\textit{but}(ga_2, p)$ also true at this index. Hence, unlike Alice's action, Beth's action is a *but-for* cause of Carl's ending up inside the shop. Similarly, since Carl would not be inside the shop if he did not try to enter, ($[cs_1ga_2gh_3]\neg p$ is true at m_1/h_1), his action is also a *but-for* cause of p ($\textit{but}(es_3, p)$ is true at m_1/h_1). Therefore, both Beth's and Carl's actions actually contributed to Carl's being inside the shop, as intuitively it should be.

So far so good. But consider a variant of Example 3.3.3.

3.3.6. Example. Everything is as in Example 3.3.3, except that there is an additional agent, Diana, who, instead of going home, forcefully pushes Carl inside the shop. Diana is so strong that she would have succeeded in pushing Carl inside, even if Beth did not grab Alice.

Let Diana be agent 4 and pc_4 and gh_4 stand for the action types *Diana's pushing Carl inside the shop* and *Diana's going home*. In this version of the story, none of

the actions that are actually performed (i.e., cs_1, ga_2, es_3, and pc_4) pass the *but-for* test for p. In fact, since Diana successfully pushes Carl inside the shop regardless of the actions of the other agents, none of the latter actions are necessary for Carl's being inside the shop. In turn, given Beth's action, Carl successfully enters the shop even if Diana did not push him, and so Diana's action is not necessary for Carl's being inside the shop either. Yet, intuitively, the actions that Beth, Carl, and Diana perform all contribute to Carl's being inside the shop.

The NESS test for actual causation [Wright, 1988, 2013] was specifically designed to overcome this difficulty. The idea is to weaken the necessity requirement encoded by the *but-for* test: the candidate cause need not be a necessary condition *for the result*; rather, it only need to be a necessary condition *for an actual event to be sufficient for the result*. In our terminology, instead of requiring that an action α_I is necessary for a certain result φ given what all other agents do, the NESS test requires that α_I is necessary for an action that includes α_I to be sufficient for φ. That is: α_I must be a sub-action of an action β_J that is sufficient for φ and such that β_J minus α_I is not sufficient for φ. It is generally accepted that the sufficient condition β_J must be *minimal*, in the sense that none of its constituents is sufficient for the result by itself [Mackie, 1965, 1974; Wright, 2013]. This leads us to the following definition:

3.3.7. Definition (NESS cause). Where $\alpha_I \in G\text{-}Acts$ and $\varphi \in \mathcal{L}_{\mathsf{ALO}_n}$,

$$ness(\alpha_I, \varphi) := \bigvee_{\substack{\beta_J \in G\text{-}Acts: \\ \alpha_I \sqsubseteq \beta_J}} (do(\beta_J) \wedge [\beta_J]\varphi \wedge \bigwedge_{K \subset J} (\neg[\beta_K]\varphi))$$

According to Definition3.3.7, α_I *is a NESS-cause of* φ just in case α_I is part of an *actual* condition ($\alpha_I \sqsubseteq \beta_J$ and $do(\beta_J)$) that is *sufficient* for φ ($[\beta_J]\varphi$) and *minimally* so ($\bigwedge_{K \subset J} \neg[\beta_K]\varphi$).

Going back to Example 3.3.6, let us (informally) check that each of ga_2, es_3, and pc_4 is a NESS cause of p. Since the group action ga_2es_3 results in Carl's being in the shop regardless of what Alice and Diana do, this action is a sufficient condition for p. On the other hand, neither ga_2 nor es_3 are sufficient conditions for p: given the circumstances, Carl would not end up inside the shop if he did not try to enter (this explains why ga_2 is not sufficient for p) or if Beth did not grab Alice (this explains why es_3 is not sufficient for p). We can then conclude that ga_2es_3 is a *minimal* sufficient condition for p. Its sub-actions ga_2 and es_3 are thus NESS causes of p. Finally, since Diana's action pc_4 results in Carl's being in the shop regardless of what the other agents do and since this action does not have any sub-action, pc_3 is a minimal sufficient condition for p, and so one of its NESS causes as well.

3.3.3 Responsibility operators

Having defined the notions of expected result, *but-for* cause, and NESS cause in $\mathcal{L}_{\mathsf{ALO}_n}$, we can now represent the second and third preliminary phases in the attribution of responsibility for a state of affairs φ (the first phase is implicit in the formal representation of a case). Specifically, for any action type α_I selected as relevant, we will check whether the following hold:

1. *Identification of the potential causes of* φ
   $do(\alpha_I) \wedge (do(\alpha_I) \boxplus\!\!\rightarrow \varphi)$
2. *Selection, among the potential causes, of the actual causes of* φ
   $do(\alpha_I) \wedge (do(\alpha_I) \boxplus\!\!\rightarrow \varphi) \wedge but(\alpha_I, \varphi)$ (*but-for version*)
   $do(\alpha_I) \wedge (do(\alpha_I) \boxplus\!\!\rightarrow \varphi) \wedge ness(\alpha_I, \varphi)$ (*NESS version*)

It is natural to associate the tests in the two phases with different levels of responsibility, provided that two further conditions are met. First, *for all tests*, $\mathsf{X}\varphi$ should not be inevitable, or settled, in the considered situation. This is the constraint characterizing deliberative STIT [Horty and Belnap, 1995] and is justified by the fact that no one would be held responsible for something that would have happened regardless of her action. Second, *for the tests for actual causation*, the agent should not be forced to do the relevant action in the considered situation – there should be an alternative action available to her. This constraint is justified by the fact that, even if her action actually caused the result, the agent would not be held responsible for the result if she was forced to act the way she did. This leads us to the introduction of the following operators for responsibility.

3.3.8. Definition (Responsibility operators). Where $I \subseteq Ag$ and $\varphi \in \mathcal{L}_{\mathsf{ALO}_n}$,

$$[I\,pres]\varphi := \bigvee_{\alpha \in Ag\text{-}Acts} (do(\alpha_I) \wedge (do(\alpha_I) \boxplus\!\!\rightarrow \varphi)) \wedge \neg\Box\mathsf{X}\varphi \wedge \mathsf{X}\varphi$$

$$[I\,sres]\varphi := \bigvee_{\alpha \in Ag\text{-}Acts} (do(\alpha_I) \wedge (do(\alpha_I) \boxplus\!\!\rightarrow \varphi) \wedge but(\alpha_I, \varphi) \wedge \neg\Box\mathsf{X}\varphi \wedge \neg\Box do(\alpha_I))$$

$$[I\,res]\varphi := \bigvee_{\alpha \in Ag\text{-}Acts} (do(\alpha_I) \wedge (do(\alpha_I) \boxplus\!\!\rightarrow \varphi) \wedge ness(\alpha_I, \varphi) \wedge \neg\Box\mathsf{X}\varphi \wedge \neg\Box do(\alpha_I))$$

We call $[I\,pres]$ the operator for *potential responsibility*, $[I\,sres]$ the operator for *strong responsibility*, and $[I\,res]$ the operator for *plain responsibility*. According to Definition 3.3.8, an individual agent or a group is *potentially responsible for* φ just in case they perform an action that is expected to result in φ, it is not settled that φ happens next, and φ happens next. They are *strongly (resp. plainly) responsible for* φ just in case it is not settled that φ happens next and they perform an action that, besides being expected to result in φ, is a *but-for* (resp. NESS) cause of φ *and* is not the only action available to them. In this way, the three definitions take into account a number of elements that are usually considered essential for the attribution of moral or legal responsibility for φ [see,

e.g., Van De Poel, 2015, Section 1.3.], that is: the contingency of φ, expressed by $\neg\Box\mathsf{X}\varphi$; the presence of a direct and expected connection between what the agent does and φ, expressed by $do(\alpha_I) \boxplus\!\!\rightarrow \varphi$ (we come back to this in Section 3.4.1); the presence of an actual causal link between what the agent does and φ, expressed by either $but(\alpha_I, \varphi)$ or $ness(\alpha_I, \varphi)$; and, finally, the freedom of the agent in acting the way she does, expressed by $\neg\Box do(\alpha_I)$.

We call the notion encoded by $[I\,sres]\varphi$ *strong* responsibility because, if $[I\,sres]\varphi$ is true, then there is an action available to group I that *would prevent the realization of* φ, given what the other agents are doing. In fact, the following proposition is a consequence of Definition 3.3.5 and the condition of independence of agents.

3.3.9. Proposition. *Where* $I \subseteq Ag$, $\gamma \in Ag\text{-}Acts$, *and* $\varphi \in \mathcal{L}_{\mathsf{ALO}_n}$, *the following is valid in the class of all* ALO_n *frames:*

$$([I\,sres]\varphi \wedge do(\gamma)) \rightarrow \bigvee_{\substack{\alpha \in Ag\text{-}Acts: \\ \alpha_I \neq \gamma_I}} \Diamond(do(\gamma_{-I}\alpha_I) \wedge [\gamma_{-I}\alpha_I]\neg\varphi)$$

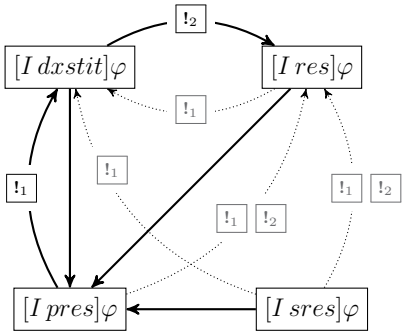
Proof:
Suppose that the antecedent it true at some index m/h of an ALO_n model $\mathcal{M}$. Then, $\mathbf{act}(m/h) = \gamma$ by the truth condition of $do(\gamma)$, and so (1) $\mathbf{act}(m/h)_I = \gamma_I$. In addition, by the def. of $[I\,sres]$, there is $\beta \in Ag\text{-}Acts$ s.t. (2) $\mathcal{M}, m/h \models do(\beta_I) \wedge but(\beta_I, \varphi) \wedge \neg\Box do(\beta_I)$. By the truth condition of $do(\beta_I)$ and (1), $\beta_I = \mathbf{act}(m/h)_I = \gamma_I$, so we can replace β_I with γ_I in (2). Since $\mathcal{M}, m/h \models \neg\Box do(\gamma_I)$, there is $h' \in H_m$ s.t. $\mathbf{act}(m/h')_I \neq \gamma_I$, by the truth condition of $\Box\varphi$ and $do(\gamma_I)$. But then there is $\alpha \in Ag\text{-}Acts$ s.t. $\alpha_I \neq \gamma_I$ and $\mathbf{act}(m/h')_I = \alpha_I$. So, for all $i \in Ag$, $\gamma_{-I}\alpha_I(i) \in Acts^m$: that is, $\gamma_{-I}\alpha_I$ is a combination of individual actions executable at m. By the condition of independence of agents, it follows that there is $h'' \in H_m$ s.t. (3) $\mathbf{act}(m/h'') = \gamma_{-I}\alpha_I$. Since $\mathcal{M}, m/h \models do(\gamma) \wedge but(\gamma_I, \varphi)$, $\mathcal{M}, m/h \models [\gamma_{-I}\beta_I]\neg\varphi$ for all $\beta \in Ag\text{-}Acts$ s.t. $\beta_I \neq \gamma_I$ by the definition of $but(\gamma_I, \varphi)$, and so $\mathcal{M}, m/h \models [\gamma_{-I}\alpha_I]\neg\varphi$. As $[\gamma_{-I}\alpha_I]\neg\varphi$ entails $\Box[\gamma_{-I}\alpha_I]\neg\varphi$, (4) $\mathcal{M}, m/h'' \models [\gamma_{-I}\alpha_I]\neg\varphi$ by the truth condition of $\Box\varphi$. It follows from (3) and (4) that $\mathcal{M}, m/h \models \Diamond(do(\gamma_{-I}\alpha_I) \wedge [\gamma_{-I}\alpha_I]\neg\varphi)$, whence the result. □

Hence, if group I is strongly responsible for φ, then I has the ability to prevent φ given what the other agents are doing. The notion encoded by $[I\,res]\varphi$ does not imply that I has such a strong negative control on the result. In fact, $[I\,res]\varphi$ does not even ensure that there is an action available to I that *might* result in $\neg\varphi$, given what the other agents are doing. In Example 3.3.6, for instance, Carl is plainly responsible for ending up inside the shop, even if, *given that Diana pushes him inside*, there is no action available to him that would (or might) prevent this result. Importantly, this does *not* mean that Carl is plainly responsible for ending up inside the shop, no matter what he does. In fact, if Carl decided to go home, his action would not be part of any actual minimal sufficient condition

An unlabeled arrow from formula A to formula C means that $A \to C$ is valid in the class of ALO_n models. A curved labeled arrow from A to C means that $A \to C$ is valid in the class of ALO_n models, *provided that* A is strengthened with additional conditions:

$\boxed{!_1}$: $\bigvee_{\alpha \in Ag\text{-}Acts} \Box(do(\alpha_I) \to \underline{do}(\alpha_I))$

$\boxed{!_2}$: $\bigvee_{\alpha \in Ag\text{-}Acts}(do(\alpha) \wedge \bigwedge_{K \subset I} \neg[\alpha_K]\varphi)$



where $I \subseteq Ag$ and $\varphi \in \mathcal{L}_{\mathsf{ALO}_n}$

Figure 3.3: Logical relations between STIT and responsibility operators

for the result, and so he would not be plainly responsible for it. More generally, the requirement that $\mathsf{X}\varphi$ is not settled true ensures that an individual agent or a group that is potentially or actually responsible for φ could have been not responsible for φ. Formally, it is not difficult to check that the three implications below are valid in the class of ALO_n models:

$$[I\,pres]\varphi \to \Diamond\neg[I\,pres]\varphi \qquad [I\,sres]\varphi \to \Diamond\neg[I\,sres]\varphi \qquad [I\,res]\varphi \to \Diamond\neg[I\,res]\varphi$$

Before moving on to the analysis of concrete examples, the diagram in Figure 3.3 schematises the logical relations between the new responsibility operators and the standard deliberative XSTIT operator (the proofs that these relations hold can be found in Appendix A.2). For clarity of exposition, let us call "STIT responsibility" the notion of responsibility encoded by the deliberative XSTIT operator. Focusing on the thick arrows, the diagram shows that potential responsibility is the logically weakest notion of responsibility among those that we have introduced, in line with the idea that strongly and plainly responsible individuals or groups are selected among potentially responsible individuals or groups. In addition, potential responsibility is also strictly weaker than STIT responsibility. This is as it should be: an agent who does something that guarantees the truth of φ does something that is expected to result in φ.

More interestingly, looking at the labeled arrow from $[I\,pres]\varphi$ to $[I\,dxstit]\varphi$, potential responsibility for φ (hence, strong and plain responsibility for φ) implies STIT responsibility for φ, *provided that* the responsible individual or group can only act unopposed. This means that, in our framework, potential responsibility coincides with STIT responsibility in those special cases in which the relevant

agents act free of any external interference. The implication

$$[I\,pres]\varphi \wedge \bigvee_{\alpha \in Ag\text{-}Acts} \Box(do(\alpha_I) \rightarrow \underline{do}(\alpha_I)) \rightarrow [I\,dsxtit]\varphi$$

thus expresses the idealization underlying standard STIT semantics that we flagged already in Section 3.1, namely that of ignoring any possible dependence between the actions of different agents or groups.

Turning to the upper part of the diagram, the labeled arrow from $[I\,dxstit]\varphi$ to $[I\,res]\varphi$ indicates that STIT responsibility for φ implies plain responsibility for φ, *provided that* the action performed by the relevant individual or group is a *minimal* sufficient condition for φ. Since individual actions always satisfy the minimality requirement, this means that *individual* STIT responsibility unconditionally implies *individual* potential responsibility. Putting this together with our previous consideration, we obtain that, if we restrict attention to situations in which there is no opposition between different actions, then individual STIT responsibility and individual plain responsibility coincide. Example 3.4.2 below will show that things are different for *group* responsibility: even when all actions can only be done unopposed, there are cases in which a group is STIT responsible but not plainly responsible for a result.

3.4 ALO_n at work

In the previous section, we saw that the responsibility operators $[I\,pres]\varphi$, $[I\,sres]\varphi$ and $[I\,res]\varphi$ bring together several important elements underlying responsibility attributions. In this section, we apply these operators to the analysis of paradigmatic examples of individual as well as group responsibility. This will allow us both to clarify how the aforementioned elements interact with one another and to provide a rigorous representation of a number of key distinctions. The main questions we aim at answering are as follows: Can the new responsibility operators be used to effectively handle cases that seem to be out of the reach of standard STIT semantics? Why does an agent's action need to be both a potential and an actual cause of a certain result for the agent to be actually responsible for that result? Isn't either one of the two conditions sufficient? Do our operators help shedding lights on the notion of group responsibility? We start by discussing cases of individual responsibility in Section 3.4.1 and move the analysis to cases of group responsibility in Section 3.4.2.

3.4.1 Individual responsibility

Let us begin by going back to our initial Example 3.1.1: Alice shoots Dan dead; Beth, a bystander, could hit Alice thus preventing Dan's death, but she remains still, petrified with fear. Figure 3.4 depicts an ALO_n model representing Example

Alice is agent 1 and can either shoot Dan (sd_1) or stand still (ss_1); Beth is agent 2 and can either hit Alice's harm (ha_2) or stand still (ss_2). q stands for the proposition that Dan is dead. sd_1 is opposed by ha_2. The gray shaded area highlights that sd_1 is unopposed in sd_1ss_2. Index m_1/h_1 represents what actually happened.

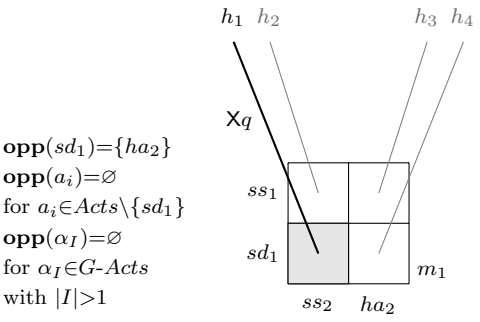


Figure 3.4: An ALO_n model representing Example 3.1.1

3.1.1. By inspecting the model and applying the definitions from Section 3.3, it is easy to check that the formulas in the following table are true at index m_1/h_1 (*amsc* means "actual minimal sufficient conditions"):

Effect:	$\mathsf{X}q$, $\neg\Box\mathsf{X}q$
Alice's action:	$do(sd_1)$, $\neg\Box do(sd_1)$, $\neg[sd_1]q$, $[ss_1]\neg q$
Beth's action:	$do(ss_2)$, $\neg\Box do(ss_2)$, $\neg[ss_2]q$, $[ha_2]\neg q$
Expected result test:	$do(sd_1) \boxplus\!\!\rightarrow q$
But-for test:	$but(sd_1, q)$, $but(ss_2, q)$
NESS test:	$ness(sd_1, q)$, $ness(ss_2, q)$ (*amsc* for q: sd_1ss_2)
Upshot for Alice:	$\neg[1\, dsxtit]q, [1\, pres]q$, $[1\, sres]q$, $[1\, res]q$
Upshot for Beth:	$\neg[2\, dsxtit]q, \neg[2\, pres]q$, $\neg[2\, sres]q$, $\neg[2\, res]q$

Hence, Alice is potentially, strongly, and plainly responsible for Dan's death, as intuitively it should be: after all, Alice did something that was expected to result in Dan's death and that, other things being equal, was both necessary and sufficient for this result. As we suggested in Section 3.1, the deliberative STIT operator cannot express this intuitive judgment. This depends on the fact that Alice's action, by itself, did not ensure Dan's death – Alice's shot would have been diverted had Beth hit Alice's arm. This is reflected, in our model, by the fact that $\neg[sd_1]q$ is true at m_1/h_1, and so $[1\, dsxtit]q$ is false at that index.

It is worth noting that, according to the proposed analysis, Alice *alone* is responsible for Dan's death: there is no sense in which Beth is causally responsible for Dan's death, even if, by not intervening, she did contribute to the result (as shown in the table, Beth's action is both a *but-for* and a NESS cause of Dan's death). This is, again, intuitive, since, after all, it was not Beth who pulled the trigger.[27] Importantly, we can capture this commonsensical conclusion because,

[27]The fact that Beth is not responsible agrees with the treatment of omissions in the Anglo-

according to our definitions, an agent can only be responsible for those states of affairs that were expected to result from her action. This suggests that the expected result test can be viewed as a test to ascertain whether there was a *substantial causal link* between the agent's action and the result or whether the agent's action contributed to the result only by chance. From this perspective, the distinction between potential and actual cause seems to be close to the distinction between legal and factual cause characterizing analyses of causation in criminal law [see Herring, 2012, Chapter 2.3]: While a *factual cause* of a fact φ is what we called an actual cause of φ (factual causation is established by applying either the *but-for* or the NESS test), a *legal cause* of φ is, roughly, something that played a *substantial role* in bringing about φ. The question of what counts as "substantial" is open-ended,[28] but, as our notion of potential cause, it aims to prune away the causal factors that are only indirectly or remotely related to the result in question.

These are all promising results, but a potential worry might arise at this point: isn't it the case that, *given what Alice does*, Beth's standing is expected to result in Dan's death? If so, then it seems that Beth is responsible for Dan's death after all.[29] Let us first emphasize that, in order to determine the expected results of the action performed by an agent i at a moment m, one needs to consider all possible ways the opponents of i could have acted at m, and not just what i's opponents did in fact. That said, in a situation in which *it is settled* that Alice shoots Dan and Beth can prevent Dan's death, Beth does indeed turn out to be causally responsible for Dan's death according to our definitions. Even worse, in this hypothetical scenario, Alice is not causally responsible for the death, because she has no choice but to shoot Dan. This result seems to be problematic, given that, as in Example 3.1.1, it is Alice and not Beth who pulls the trigger.

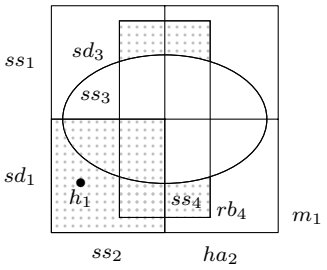
Our reply is that the story is too underspecified to determine whether the above-mentioned conclusions are counterintuitive. Concerning Alice, a relevant detail in this context is how she ended up in a situation in which shooting Dan is her only option. Presumably, in order to fill in this detail we need to consider what happened *before* Alice shot Dan. But then the present example involves agents acting over time, not just simultaneously, which, in turn, calls for a generalization of the notions of potential and actual cause to *courses of action*. We leave this generalization to future work, partly because we think that it requires a proper representation of counterfactual reasoning in STIT (we come back to this in Section 3.5). Regarding Beth, we think that, in the hypothetical scenario

American criminal law tradition, according to which the defendant can be criminally liable for a failure to act only if she is either under a duty of care or under a duty to neutralize a danger she created in the first place [see Carr and Johnson, 2013, Chapter 2]. With respect to Example 3.1.1, this means that Beth is criminally liable for Dan's death if she is, for instance, a police officer on duty, but not if she is simply a bystander (as we assume her to be).

[28] See Honoré and Gardner [2010] for a summary of the debate.

[29] We thank Masayuki Tashiro for raising this issue.

With respect to Figure 3.4, we added the following details. Carl is agent 3 and can either shoot Dan (sd_3, area outside the ellipse) or stand still (ss_3, area inside the ellipse). Diana is agent 4 and can either run between Carl and Dan (rb_4, area outside the inner rectangle) or stand still (ss_4, area inside the inner rectangle). sd_3 is opposed by rb_4. $\mathsf{X}q$ is true at m_1 on all histories passing through the dotted areas (which are omitted for the sake of readability). The black dot in the bottom left corner stands for the actual history h_1.



opp(sd_1)=$\{ha_2\}$
opp(sd_3)=$\{rb_4\}$
opp(a_i)=$\varnothing$ for $a_i \in Acts \setminus \{sd_1\}$
opp(α_I)=$\varnothing$ for $\alpha_I \in G\text{-}Acts$ with $|I|>1$

Figure 3.5: An ALO_n model representing Example 3.4.1

under consideration, she may well be *causally* responsible for Dan's death. Importantly, this, by itself, does not mean that she is *blameworthy* or *criminally liable* for it: as we mentioned at the beginning of this chapter, in order to establish blameworthiness or liability, we would need to consider Beth's intentions and beliefs. This falls outside the scope of the present investigation.

Before proceeding, let us consider an elaborated version of Example 3.1.1.

3.4.1. Example. Everything is as in Example 3.1.1, except that Carl also fires at Dan. Yet, Diana runs between Carl and Dan and Carl's bullet hits her.

Example 3.4.1 can be modeled as shown in Figure 3.5. It is not difficult to see that the formulas in the following table are true at index m_1/h_1:

Effect:	$\mathsf{X}q$, $\neg\Box\mathsf{X}q$
Alice's action:	$do(sd_1)$, $\neg\Box do(sd_1)$, $\neg[sd_1]q$, $\langle ss_1\rangle\,\neg q$
Beth's action:	$do(ss_2)$, $\neg\Box do(ss_2)$, $\neg[ss_2]q$, $\langle ha_2\rangle\, q$
Carl's action:	$do(sd_3)$, $\neg\Box do(sd_3)$, $\neg[sd_3]q$, $\langle ss_3\rangle\, q$
Diana's action:	$do(rb_4)$, $\neg\Box do(rb_4)$, $\neg[rb_4]q$, $\langle ss_4\rangle\, q$
Expected result test:	$do(sd_1) \boxplus\!\!\rightarrow q$, $do(sd_3) \boxplus\!\!\rightarrow q$
But-for test:	$but(sd_1, q)$, $but(ss_2, q)$
NESS text:	$ness(sd_1, q)$, $ness(ss_2, q)$ ($amsc$ for q: $sd_1 ss_2$)
Upshot for Alice:	as in Example 3.1.1
Upshot for Beth:	as in Example 3.1.1
Upshot for Carl:	$\neg[3\,dsxtit]q$, $[3\,pres]q$, $\neg[3\,sres]q$, $\neg[3\,res]q$
Upshot for Diana:	$\neg[4\,dsxtit]q$, $\neg[4\,pres]q$, $\neg[4\,sres]q$, $\neg[4\,res]q$

Hence, in this case, both Alice and Carl are potentially responsible for Dan's death, and rightly so: since they did something that was expected to result in

Dan's death, we would surely include both of them in our list of candidates for actual responsibility. Still, only Alice's action was necessary and sufficient for Dan's death given the circumstances, and so only Alice is actually responsible for it. This agrees with the fact that, in a case like this, Carl could be accused, at most, of attempted murder but surely not of homicide.

3.4.2 Group responsibility

So far we have considered cases in which one of the actions of the relevant agents was both necessary and sufficient for a certain result, other things being equal. The most interesting cases of group responsibility are those in which the actions of multiple agents were either separately sufficient or jointly necessary for the result. Let us consider two paradigmatic examples.

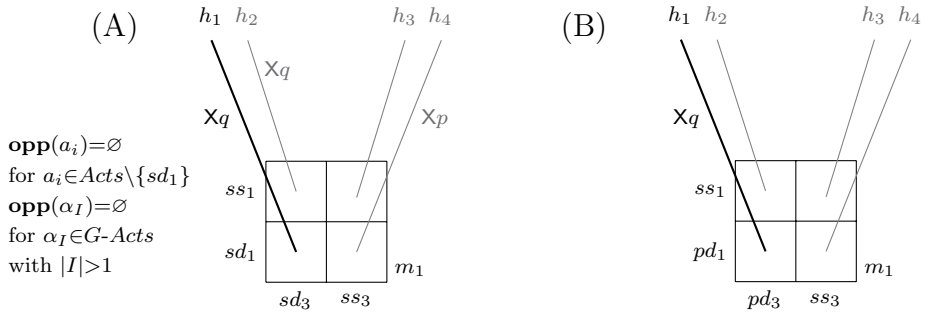
3.4.2. Example. Alice and Carl simultaneously fire at Dan and shoot him dead. Each shot, alone, would have been sufficient to kill Dan.

Figure 3.6(A) shows how Example 3.4.2 can be modeled in our framework. The analysis of the case is summarized in the following table, where, as before, the listed formulas are true at index m_1/h_1.

Effect:	$\mathsf{X}q$, $\neg\Box\mathsf{X}q$
Alice's action:	$do(sd_1)$, $\neg\Box do(sd_1)$, $[sd_1]q$, $\langle ss_1\rangle\, q$
Carl's action:	$do(sd_3)$, $\neg\Box do(sd_3)$, $[sd_3]q$, $\langle ss_3\rangle\, q$
Expected result test:	$do(sd_1) \boxplus\!\!\rightarrow q$, $do(sd_3) \boxplus\!\!\rightarrow q$, $do(sd_1 sd_3) \boxplus\!\!\rightarrow q$
But-for test:	none
NESS text:	$ness(sd_1, q)$, $ness(sd_3, q)$ (*amsc* for q: sd_1 and sd_3)
Upshot for Alice:	$[1\, dsxtit]q, [1\, pres]q$, $\neg[1\, sres]q$, $[1\, res]q$
Upshot for Carl:	$[3\, dsxtit]q, [3\, pres]q$, $\neg[3\, sres]q$, $[3\, res]q$
Upshot for *Ag***:**	$[Ag\, dsxtit]q, [Ag\, pres]q$, $\neg[Ag\, sres]q$, $\neg[Ag\, res]q$

As it should be expected from our discussion in Section 3.3.2, in this case both Alice and Carl are potentially and plainly responsible for Dan's death, but neither of them is strongly so. The reason is that neither Alice's nor Carl's action was necessary, in the circumstances, for Dan's death.

More interestingly, according to our analysis, the group consisting of Alice and Carl is potentially but neither strongly nor plainly responsible for the result. The failure of strong responsibility depends on the fact that the execution of the joint action $sd_1 sd_3$ was not necessary for Dan's death: Dan would have died even if $sd_1 ss_3$ or $ss_1 sd_3$ had been executed. The failure of plain responsibility depends on the fact that $sd_1 sd_3$ was not a *minimal* sufficient condition of Dan's death. So, according to our analysis, the group of Alice and Carl does not pass any test for actual responsibility – even if it is a candidate for it. If we analyze the



Alice is agent 1 and Carl is agent 3. In Figure (A) each agent $i \in \{1, 3\}$ can either shoot Dan (sd_i) or stand still (ss_i), while in Figure (B) each agent $i \in \{1, 3\}$ can either push Dan (pd_i) or stand still (ss_i). No action is opposed. q stands for the proposition that Dan is dead. In both figures, the actual index is m_1/h_1.

Figure 3.6: ALO_n models representing Examples 3.4.2 and 3.4.3

example using the deliberative STIT operator, we obtain a different result: since $[sd_1]q$, $[sd_3]q$, and $[sd_1 sd_3]q$ are true at m_1/h_1, all of $[\{1\}\, dsxtit]q$, $[\{3\}\, dsxtit]q$, and $[\{1, 3\}\, dsxtit]q$ are true at this index. We will come back to this in a moment.

3.4.3. Example. Alice and Carl push Dan, who falls down a cliff and dies on impact. Neither push, by itself, would have made Dan fall.

Example 3.4.3 can be modeled as shown in Figure 3.6(B). Again, it is not difficult to see that the following formulas are true at m_1/h_1.

Effect:	$\mathsf{X}q$, $\neg\Box\mathsf{X}q$
Alice's action:	$do(pd_1)$, $\neg\Box do(pd_1)$, $\neg[pd_1]q$, $[ss_1]\neg q$
Carl's action:	$do(pd_3)$, $\neg\Box do(pd_3)$, $\neg[pd_3]q$, $[ss_3]\neg q$
Expected result test:	$do(pd_1 pd_3) \boxplus\!\!\rightarrow q$
But-for test:	$but(pd_1, q)$, $but(pd_3, q)$, $but(pd_1 pd_3, q)$
NESS test:	$ness(pd_1, q)$, $ness(pd_3, q)$, $ness(pd_1 pd_3, q)$
	(*amsc* for q: $pd_1 pd_3$)
Upshot for Alice:	$\neg[1\, dsxtit]q, \neg[1\, pres]q$, $\neg[1\, sres]q$, $\neg[1\, res]q$
Upshot for Carl:	$\neg[3\, dsxtit]q, \neg[3\, pres]q$, $\neg[3\, sres]q$, $\neg[3\, res]q$
Upshot for *Ag*:	$[Ag\, dsxtit]q, [Ag\, pres]q$, $[Ag\, sres]q$, $[Ag\, res]q$

In this case, neither Alice nor Carl is individually responsible for Dan's death in any of the senses we introduced, although each of their actions is a *but-for* as well as a NESS cause of it. The reason is that neither Alice's nor Carl's

action was expected to kill Dan. By contrast, the group of Alice and Carl is both potentially and actually responsible for the death. The reason is that the action pd_1pd_3 performed by the group was expected to kill Dan and its execution was both necessary and minimally sufficient for this result. The analysis in terms of STIT operators gives the same result in this case: on the one hand, since $\neg[pd_1]q$ and $\neg[pd_3]q$ are true at m_1/h_1, both of $[1\,dsxtit]q$ and $[3\,dsxtit]q$ are false at this index; on the other hand, since $[pd_1pd_3]q$ is true at m_1/h_1, $[\{1,3\}\,dsxtit]q$ is also true at this index.

The upshot is that, while according to our theory the group of Alice and Carl is actually responsible for Dan's death in Example 3.4.3 but not in Example 3.4.2, according to STIT the group is equally responsible in the two cases. However, the members of the group are individually responsible only in Example 3.4.2 according to both our theory and STIT.

How should we read these results? Let us go back to individual responsibility before lifting the discussion to groups. As noted by a number of moral philosophers [see, e.g., Bernstein, 2017; Sartorio, 2015], in cases like Example 3.4.2 and Example 3.4.3 it is intuitive to argue in two opposite directions. Let us focus on Alice (the analysis for Carl is analogous). On the one hand, we can say that in Example 3.4.2 Alice is more responsible than in Example 3.4.3 because, by acting the way she did, she ensured Dan's death in the former but not in the latter case: in Example 3.4.2 Dan's death was entirely Alice's fault. On the other hand, we can say that in Example 3.4.3 Alice was more responsible than in Example 3.4.2 because, by acting differently, she could have prevented Dan's death in the former but not in the latter case: in Example 3.4.2, Dan would have died regardless of Alice's shot, and so Dan's death was, in a sense, independent of her choice.

Interestingly, our proposal breaks the symmetry between the two arguments by adding a further parameter, namely the presence of a substantial causal link between the actions in question and the result. According to our analysis, Alice is *less* responsible in Example 3.4.3 than in Example 3.4.2, because in the former case her action was not expected to result in Dan's death. This means that we cannot exclude that her contribution was only accidental. In contrast, in Example 3.4.2, by shooting at Dan, Alice did something that was indeed supposed to kill him. So, we can exclude that her contribution was merely accidental in this case.

The verdict is reversed for the group of Alice and Carl. The group is *more* responsible in Example 3.4.3 because in this case the result cannot be traced back to either one of the two agents. In Example 3.4.2, on the other hand, there is no residual responsibility for the group because Dan was killed by the shot of each of the two agents.

Highlighting these asymmetries is important. It shows that, in our framework, group responsibility is only ascribed when there is no other option, that is, when no member of the group can be singled out as a substantial contributor to the result. In philosophy, the problem of determining the degree of responsibility of each member of a group in cases of this sort is known as the *problem of many*

hands [Thompson, 1980]. What we are suggesting is that, in our framework, there is an intuitive correspondence between cases of group responsibility and cases in which the problem of many hands arises. In this regard, it is interesting to consider one last example.

3.4.4. Example. Alice, Beth, and Carl push Dan, who falls down a cliff and dies on impact. Neither push, by itself, would have made Dan fall but two pushes would have been sufficient.

An ALO_n model representing Example 3.4.4 can be obtained by modifying the model in Figure 3.6(B) in the obvious way. It is not difficult to see that, according to our definitions, neither Alice nor Beth nor Carl is going to be individually responsible for Dan's death, in any of the senses we introduced. In addition, although potentially responsible for the death, the group of the three agents is going to be neither strongly nor plainly responsible for it: since Dan would have died even if one of the three agents had not pushed him, the action of the group is neither necessary nor *minimally* sufficient for Dan's death. On the other hand, each group of two agents is going to be potentially as well as plainly responsible for Dan's death (the lack of strong responsibility depends on the fact that the example involves overdetermination). We are thus going to obtain an intuitive "medium level" of responsibility: no individual agent is going to be responsible; every group of two agents is going to be responsible; the group of all agents is not going to be responsible. If the actions of all of Alice, Beth, and Carl were necessary for Dan's death, the upshot for the groups would be reversed, in line with the intuition that responsibility would be distributed in different ways in the two cases.[30]

3.5 Conclusion

In this chapter, we presented a refinement of STIT semantics for reasoning about causal responsibility when the involved agents can interfere with one another. We formalized three tests to ascribe causal responsibility, namely the expected-result test for potential causality and the *but-for* and NESS tests for actual causality. We used these tests to define operators for corresponding levels of causal responsibility and argued that the new operators deliver promising results in the analysis of paradigmatic examples. Specifically: (1) they allow us to handle cases of individual responsibility that are out of the reach of standard STIT semantics, like Example 3.1.1; (2) they are sensitive to the distinction between agents who substantially contribute to a result and agents who contribute to a result but not substantially, like, respectively, Alice and Beth in Example 3.1.1; (3) they are

[30] See Kaiserman [2018] for a concise overview of the main measures of degrees of responsibility that have been discussed in the literature.

sensitive to the distinction between agents who actually contribute to a result and agents who only attempt to bring it about, like, respectively, Alice and Carl in Example 3.4.1; (4) unlike STIT operators, they allow us to highlight important asymmetries in cases of group responsibility in which the same group contributes to a result in different ways, like the group of Alice and Carl in Examples 3.4.2 and 3.4.3; (5) they allow us to identify an intuitive "medium level" of responsibility in cases in which the actions of some but not all agents are necessary for a result, like Example 3.4.4. We think that a virtue of our proposal is that we achieved these results in a relatively simple way, that is, by only supplementing STIT with action types and a basic relation of opposing between them.

We came across several open questions in the course of this chapter. First, given the central role of the relation of opposing, it is natural to ask whether this relation can be further analyzed, e.g., in terms of goals of actions, as suggested in footnote 11. Including a representation of goals seems to be important not only to clarify what the opposing relation is but also to account for the fact that an agent is not responsible for any result of her actions whatsoever. For instance, suppose that the sound of Alice's shot is sufficient to make Beth faint. Then, Alice is responsible for Beth's fainting according to our definitions, even though this is not intuitively so. A way to explain this intuition is that Beth's fainting is not intrinsically related to the goal of Alice's action (say, injuring Dan). This could be made precise by resorting to a theory of content-preserving entailment, like truthmaker semantics [see Fine, 2017] or a topic sensitive semantics for intentional modals [see Berto, 2018; Chapter 5].

A second open question is how to generalize the notions of expected result and actual cause in order to cover cases in which the relevant agents act over time, rather than just "in one step." As we mentioned in Section 3.4.1, we think that this generalization requires a proper representation of counterfactual reasoning in STIT. For instance, suppose that Beth plants a device in Alice's brain at moment $m_0 < m_1$ that forces Alice to pull the trigger at m_1. Then, in order to determine whether Alice is responsible for Dan's death, we should answer the following question: Would Alice have pulled the trigger had Beth not planted the device beforehand? Questions like the latter involve counterfactuals about agents acting over time. There are two issues related to such counterfactuals that are important for the formulation and assessment of responsibility judgments: (1) *When are these counterfactuals true?*, and (2) *How does the cognitive process that we use to evaluate them work? What is its epistemic value?*

Elaborating on the framework introduced in this chapter, in the next Chapter 4 we will address issue (1) by providing a STIT *semantics* for what we will call choice-driven counterfactuals. We will gradually introduce three new candidate semantics and discuss their logical as well as philosophical implications, connecting them to issues from the game- and decision-theory literature. In the chapter after that, we will turn to issue (2) and investigate the *cognitive tool* that we use to evaluate counterfactuals, namely imagination as reality-oriented mental

simulation – the activity of simulating hypothetical scenarios in one's mind and explore what would happen if they were realized. By drawing on research in cognitive psychology and the philosophy of mind, Chapter 5 examines the structure and logic of this activity and considers the key question how we can gain knowledge via it.

Chapter 4

STIT semantics for choice-driven counterfactuals

What would have happened if the charge nurse had not put the wrong medications on the desk? Would the intern have given them to the patient anyway? What if Zac hadn't moved out of the way? Would the thief have shot him? Would Beth's husband have picked up the kids if she hadn't? If David had bet tails, would Max have left the game? These types of questions play a central role in many situations, including when we determine responsibility, when we make plans for the future, and when we reason strategically about how our choices influence the choices of others. A common feature of these questions is that they involve what we will call *choice-driven counterfactuals*. Choice-driven counterfactuals are counterfactuals whose semantic value depends on how agents are expected to act. This means that the evaluation of a choice-driven counterfactual relies on auxiliary premises about the *default choice behavior* of the involved agents, where the default choice behavior is determined by, for instance, duties, personality, daily schedule, preferences, goals, and so on.

Our aim in this chapter is to study the semantics and logic of choice-driven counterfactuals. To do this, we improve the STIT semantics introduced in Chapter 3 in order to represent the past and future default choice behavior of the agents and to refer to moments and histories that, although no longer possible, would still be possible had something different happened in the past. We show how to merge the new framework with the mainstream semantics of counterfactuals due to Stalnaker [1968] and Lewis [1973a], highlighting important philosophical issues and interesting logical properties of choice-driven counterfactuals.

Since counterfactual reasoning is key to a number of applications of STIT (such as the analysis of the notion of responsibility, as we saw in Chapter 3), it would not be surprising if the question how to interpret counterfactuals in STIT semantics had already been addressed in the literature. As far as we know, however, only Xu [1997] and Horty [2001, Chapter 4] explicitly raise this question. In addition, although there has been some investigation concerning the semantics

of counterfactuals in the context of branching time [Placek and Müller, 2007; Thomason and Gupta, 1981], the notion of agency is absent from these proposals. One of the main contributions of the work presented in this chapter is to begin to fill this important gap in the STIT literature.

Outline. In Section 4.1, we introduce and motivate the semantic ingredients of our framework. A key component of our semantics is to distinguish between *deviant* and *non-deviant* actions at a moment. As explained below, an action available to an agent is deviant if it does not match the agent's *default choice behavior*. In Section 4.2, we present the syntax [Section 4.2.1] and semantics [Section 4.2.2] of our STIT logic with action types and deviant action (ALD_n), discuss a potential axiomatization [Section 4.2.2] and draw connections with strategic and epistemic STIT [Section 4.2.3]. Section 4.3 extends the logic from Section 4.2 to include counterfactuals. In Section 4.3.1, we gradually introduce two candidate semantics for choice-driven counterfactuals, one called *rewind models* inspired by Lewis [1979] and the other called *independence models* motivated by well-known counterexamples to Lewis' proposal [Slote, 1978]. The logical properties of the two semantics are studied in Section 4.3.2. Taking a cue from the literature on epistemic game theory [Stalnaker, 1996; Halpern, 2001], in Section 4.4, we consider how to evaluate choice-driven counterfactuals at moments arrived at by some agents performing a deviant action. Section 4.5 concludes.

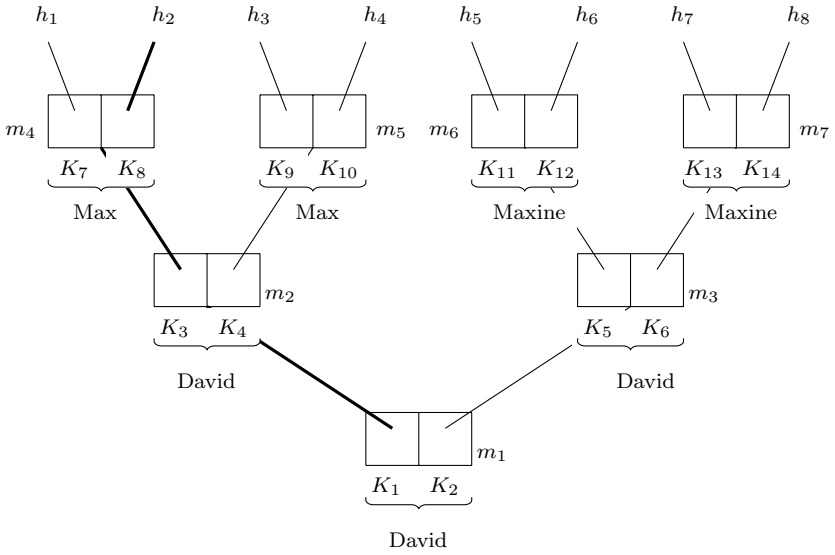
This chapter is based on Canavotto and Pacuit [2020].

4.1 Counterfactuals, agency, and branching time

Our aim in this chapter is to use STIT models to provide a semantics for choice-driven counterfactuals in branching time. Let us motivate our extensions to standard STIT semantics by gradually unrolling the following example.

4.1.1. Example. Three agents play the following game: Initially, David decides whether to play with Max or Maxine and then he bets heads or tails. After David bets, the person nominated by David flips a coin. David wins if his bet matches the outcome of the coin flip and loses otherwise. Unknown to David, both Max and Maxine have two coins, one with heads on each side and one with tails on each side (called the *H-coin* and *T-coin*, respectively). Max can see the coin he flips, and chooses so as not to match David's bet: if Max has a chance to play, he flips the H-coin if David bets tails and the T-coin if David bets heads. Maxine cannot see the coin she flips, so, if she has a chance to play, she flips a coin at random. After selecting Max, David bets heads and Max flips the T-coin.

A STIT model representing Example 4.1.1 is pictured in Figure 4.1, where



At m_1 David selects either Max (K_1) or Maxine (K_2); then, he either bets heads (K_3, K_5) or tails (K_4, K_6). If Max is selected, he either flips the H-coin (K_7, K_9) or the T-coin (K_8, K_{10}). If Maxine is selected, she either flips the H-coin (K_{11}, K_{13}) or the T-coin (K_{12}, K_{14}). The actual history is h_2.

Figure 4.1: Example 4.1.1 in standard STIT semantics

the actual history is h_2 (the thick line).[1] Suppose that we are at moment m_4 on history h_2 (so, David and Max have made their choices). Intuitively, the following counterfactual, abbreviated as $T \,\Box\!\!\rightarrow L$, is true at moment m_4 on history h_2:

($C1$) If David had bet tails, then he would still have lost.

In order to evaluate ($C1$), we need to consider the histories on which the antecedent T is true (called *T-histories*). The T-histories are the histories on which David bets tails. That is, the histories on which David performs an action of type *betting tails*. According to the explanation of Figure 4.1, David bets tails on histories h_3, h_4, h_7, and h_8. But, despite our informal description of the diagram, in STIT semantics there is nothing that allows us to group together K_4 and K_6 as actions of type *betting tails* – and so to identify the aforementioned histories as histories on which David performs this type of action. Continuing in the vein of Chapter 3, we will obviate this difficulty by supplementing our models with labels tagging the actions available to every agent at a moment with their *types*.

[1] We assume that the agents who do not move at a moment m only have one available choice at m – i.e., the *vacuous choice* represented by the set of all histories passing through m. For the sake of readability, we have omitted vacuous choices from Figure 4.1.

However, introducing action types alone is not sufficient to identify the T-histories needed to evaluate $(C1)$. The first problem is that in order to evaluate $(C1)$, we need to consider histories on which David bet tails just previous to *the time of* m_4 (the time of utterance). Histories on which he did not just bet tails but did bet tails, say, two weeks ago or will bet tails four days from now are immaterial. Given our informal description of the model in Figure 4.1, moments m_4, m_5, m_6 and m_7 are the only ones that could occur at the time that $(C1)$ is evaluated. We will use *instants* from Belnap et al. [2001] to group together moments occurring at the same time.

With the addition of types and instants, we can formally identify h_3, h_4, h_7, and h_8 as the histories on which David just bet tails. But not all of these histories are *relevant* to evaluate $(C1)$. We can ignore histories h_7 and h_8: On these histories David just bet tails at the time of m_4 *after nominating Maxine instead of Max*. This is a much bigger difference from the actual history h_2 than h_3 and h_4, on which David nominates Max (as he actually does). So, the usual considerations of minimal change [see Section 4.3] suggest focusing on h_3 and h_4. But there is also a crucial difference between h_3 and h_4. On both histories, David just bet tails at the time of m_4 after nominating Max. Yet, after that, Max flips the H-coin on h_3 and the T-coin on h_4. The key difference is that only h_3 is consistent with Max's default choice behavior, namely that *if he has a chance to play, he flips the coin that makes David lose*. Thus, we take $(C1)$ to be true assuming that Max's choice matches his default choice behavior. Contrast $(C1)$ with the counterfactual: "If David had nominated Maxine and bet tails, then he would still have lost." Given that Maxine might choose to flip the T-coin, this counterfactual is judged false.[2]

In order to represent the default choice behavior of the agents over time, we will supplement STIT semantics with a function that identifies the *deviant actions* at each moment. An action available to an agent i at a moment m is deviant if its performance at m does not agree with agent i's default choice behavior at m – it is a *non-deviant* or *default action* otherwise. To simplify the exposition, we call an agent's default choice behavior a *choice rule*. In Example 4.1.1, "Max flips the coin that makes David lose" is a choice rule and actions K_7 (flipping the H-coin after David bets heads) and K_{10} (flipping the T-coin after David bets tails) are deviant actions. We conclude this section with three clarifying comments about choice rules.

What choice rules are (not). Choice rules can have various sources, including social conventions, shared standards of rationality, habits, individual preferences or goals, and, in the case of artificial agents, choice-guiding programs. A natural example of choice rules are the *decision rules* found in the game- and decision-

[2]What is intuitively true is the weaker "If David had nominated Maxine and bet tails, he *might* have lost."

theory literature, such as *expected utility maximization* or *maximin*. However, it is important to stress that some choice rules can be dictated by habits or behavior that is, on the face of it, irrational (more on this in Section 4.4). A final point about the interpretation of choice rules is that they should *not* be thought of as *physical or causal laws*. The key difference is that the latter laws constrain the behavior of the agents in a way that choice rules do not: while an agent who is hit on his legs by a 100kg rolling ball cannot avoid falling, an agent who normally cheats at cards can avoid cheating.

(Non-)deterministic choice rules. A choice rule is *deterministic* when, at every moment m at which it guides the behavior of an agent i, there is only one available action for i that is non-deviant. Otherwise, a choice rule is said to be *non-deterministic*. Max's choice rule in Example 4.1.1 is an example of a deterministic choice rule: provided that Max can play, flipping the T-coin is his only non-deviant option if David bets heads and flipping the H-coin is his only non-deviant option if David bets tails. An example of non-deterministic choice rule is: "if mango, pineapple, and pear are available, then Alice picks either mango or pineapple." This rule guides Alice's behavior when all three fruits are present, since picking the pear is deviant. But the guidance is only partial since picking the mango and picking the pineapple are both non-deviant. Here we make the simplifying assumption that all choice rules are deterministic. Restricting to deterministic choice rules simplifies our formal definitions. Of course, this is a significant assumption since non-deterministic choice rules are ubiquitous. However, the issues concerning choice-driven counterfactuals addressed here do not depend on this assumption.

Extensional perspective on choice rules. Our models represent the distinction between actions that are deviant and actions that are not deviant according to an underlying set of choice rules. But we do not include a representation of the underlying choice rules themselves.[3] Using this approach, we can represent a wide variety of choice rules, including choice rules that may change over time. For example, we can easily represent the choice rule "Alice normally cheats at cards up to time `t` and normally respects the rules afterwards" by classifying all instances of Alice's non-cheating up to `t` as deviant and all instances of Alice's cheating after `t` as deviant. Similarly, we can represent choice rules such as "Alice is indifferent between mango and pineapple but strictly prefers watermelon over mango and pineapple": according to this rule, picking watermelon is the only non-deviant option for Alice when watermelon is available, while none of her options is deviant at moments when watermelon is not available.

[3]For instance, one could make choice rules explicit using default logic as in Horty [2012]. We leave an exploration of this possibility to future work.

4.2 Basic framework

In this section, we improve the framework introduced in Chapter 3 in order to be able to describe past facts, properties of instants, and deviant actions. The logic ALD_n we are about to present includes the fragment of the logic ALO_n from Chapter 3 *without the relation of opposing between action types*. The formal language is defined in Section 4.2.1 and the semantics is presented in Section 4.2.2, where we also consider a potential axiomatization. In Section 4.2.3 we briefly discuss connections with epistemic and strategic STIT. For the reader's convenience, we do not assume familiarity with the notation introduced in Chapter 3. This means that, although the presentation will be more concise in several parts, there will be some overlap between the present section and Chapter 3.2.

4.2.1 Syntax

We start by fixing sets of atomic propositions, action types and agents:

- Let $Prop$ be a non-empty countable set of propositional variables.
  (We will use $p, q, r, p', p'', \ldots$ for elements of $Prop$.)
- Let Atm be a non-empty finite set of (names of) action types.
  (We will use a, b, c, $a', a'', \ldots$ for elements of Atm.)
- Let $Ag = \{1, \ldots, n\}$ be the set of n agents for some number $n \in \mathbb{N}$.
  (We use will $i, j, k, i', i'', \ldots$ for elements of Ag.)

We think of agents as endowed with a repertoire of action types of which they can be authors. We associate each action type with its possible authors and fix a set $Acts$ of *individual action types*, defined as follows:

$$Acts \subseteq Atm \times Ag$$

We write a_i when $(a, i) \in Acts$. Intuitively, a_i is the action type that is instantiated whenever agent i performs an action of type a. For instance, if $a \in Atm$ is the action type *betting tails* and $1, 2 \in Ag$ are, respectively, David and Max, then a_1 is the action type *David's betting tails* and a_2 is the action type *Max's betting tails*. For $i \in Ag$, $Acts_i$ is the set of action types authored by agent i:

$$Acts_i = \{a_i \in Acts \,|\, a \in Atm\}.$$

4.2.1. Definition (Syntax of $\mathcal{L}_{\mathsf{ALD}_n}$). Let $Prop$, Atm and Ag be defined as above. The set of formulas of the language $\mathcal{L}_{\mathsf{ALD}_n}$, also denoted with $\mathcal{L}_{\mathsf{ALD}_n}$, is generated by the following grammar:

$$\varphi := p \mid do(a_i) \mid dev(a_i) \mid \neg\varphi \mid (\varphi \wedge \varphi) \mid \Box\varphi \mid \mathsf{X}\varphi \mid \mathsf{Y}\varphi$$

where $p \in Prop$ and $a_i \in Acts$. The abbreviations for the Boolean connectives are defined as usual. We use $\Diamond\varphi$, $\hat{\mathsf{X}}\varphi$, and $\hat{\mathsf{Y}}\varphi$ as abbreviations for $\neg\Box\neg\varphi$, $\neg\mathsf{X}\neg\varphi$, and $\neg\mathsf{Y}\neg\varphi$ respectively. We adopt the usual rules for the elimination of parentheses.

The three modalities of $\mathcal{L}_{\mathsf{ALD}_n}$ are standard in branching time logic: $\Box\varphi$ means "φ is settled true" or "φ is historically necessary," $\mathsf{X}\varphi$ means "φ is true at the next moment on the current history," and $\mathsf{Y}\varphi$ means "φ is true at the previous moment on the current history." The intended interpretations of the action formulas $do(a_i)$ and $dev(a_i)$ are "agent i does action a" and "action a_i is deviant," respectively. Modalities in the spirit of PDL and STIT can be defined in $\mathcal{L}_{\mathsf{ALD}_n}$ as they are defined in $\mathcal{L}_{\mathsf{ALO}_n}$ [cf. Definition 3.2.5].

We will use the following notions later in the chapter. A *complete group action* is a function $\alpha : Ag \to Acts$ such that, for all $i \in Ag, \alpha(i) \in Acts_i$. So, a complete group action is any combination of individual actions, one for each agent (in game-theoretic terms, it is an *action profile*). Let *Ag-Acts* be the set of all complete group actions (we use Greek letters α, β, γ for elements of *Ag-Acts*). As usual, when $\alpha \in$ *Ag-Acts* and $I \subseteq Ag$, we will write α_I for the restriction of α to the set I, α_{-I} for $\alpha_{Ag \setminus I}$, and $\alpha(I)$ for the image of I under α. For any $\alpha \in$ *Ag-Acts*, we define:

$$do(\alpha) := \bigwedge_{a_i \in \alpha(Ag)} do(a_i).$$

Thus, $do(\alpha)$ means "the group of all agents does α" (i.e., "for all $i \in Ag$, i performs action $\alpha(i)$").

4.2.2 Semantics

The semantics for formulas from $\mathcal{L}_{\mathsf{ALD}_n}$ consists of three main components: a *rooted discrete branching time structure* with instants (called rooted DBT structure); an *action type function* labeling moment-history pairs with complete group actions; finally, a *deviant-choice function* representing the agents' default choice behavior.

4.2.2. Definition (Rooted DBT structure). A rooted DBT structure is a tuple $\langle Mom, m_0, < \rangle$, where $Mom \neq \varnothing$ is a set of moments, $m_0 \in Mom$, and $< \subseteq Mom \times Mom$ is the temporal precedence relation. As usual, $\leq \subseteq Mom \times Mom$ is defined as: for any $m, m' \in Mom$, $m \leq m'$ iff $m < m'$ or $m = m'$. The relation $<$ is a discrete tree-like ordering of Mom rooted in m_0: it satisfies, for all $m, m_1, m_2, m_3 \in Mom$,

1. *Irreflexivity*: $m \not< m$.
2. *Transitivity*: if $m_1 < m_2$ and $m_2 < m_3$, then $m_1 < m_3$.
3. *Past-linearity*: if $m_1 \leq m_3$ and $m_2 \leq m_3$, then either $m_1 \leq m_2$ or $m_2 \leq m_1$.

(I) Histories

- A *history* is a maximal set of linearly ordered moments from Mom.
- $Hist^{\mathcal{T}}$ is the set of histories in $\mathcal{T}$.
- History h *passes through moment* m when $m \in h$.
- $H_m^{\mathcal{T}} = \{h \in Hist^{\mathcal{T}} \mid m \in h\}$ is the set of histories passing through m.
- $h, h' \in Hist^{\mathcal{T}}$ are *undivided* at m iff $m \in h \cap h'$ and there is $m' > m$ s.t. $m' \in h \cap h'$.

(II) Immediate successors

- $succ(m) = \{m' \in Mom \mid m < m'$ and, for no $m'' \in Mom, m < m'' < m'\}$ is the set of *immediate successors* of m.
- If $h \in H_m^{\mathcal{T}}$, the *immediate successor of* m *on* h, denoted with $succ_h(m)$, is the unique element of $h \cap succ(m)$.

(III) Predecessors

- If $m \neq m_0$, $pred(m)$ is the unique *predecessor* of m.

(IV) Indices

- An *index* is a pair m/h such that $m \in Mom$ and $h \in H_m^{\mathcal{T}}$.
- $Ind^{\mathcal{T}}$ is the set of indices in $\mathcal{T}$.

Table 4.1: Key notions related to a rooted DBT structure $\mathcal{T}$

4. *Discreteness*: if $m_1 < m_2$, then there is an m_3 such that $m_1 < m_3 \leq m_2$ and there is no m_4 such that $m_1 < m_4 < m_3$.

5. *No endpoints*: there is an $m' \in Mom$ such that $m < m'$.

6. *Initial moment*: $m_0 < m$.

The standard notions used to reason about rooted DBT structures are summarized in Table 4.1. We already discussed the notions in groups (I) and (IV) in Chapter 2.1 and the notions in group (II) in Chapter 3.2.1. For the remaining group, the condition of past-linearity ensures that every non-initial moment $m \neq m_0$ has a unique predecessor, denoted $pred(m)$. As usual, we will omit the superscript $\mathcal{T}$ and write $Hist$, H_m, and Ind when it is clear from the context.

We now supplement rooted DBT structures with instants. Intuitively, an instant is a set of moments happening at the same time.

4.2.3. Definition (Instants). Let $\mathcal{T} = \langle Mom, m_0, < \rangle$ be a rooted DBT structure. For any $m \in Mom$ and $n \in \mathbb{N}$, define $succ^n(m)$ recursively as follows:

1. $succ^0(m) = \{m\}$;

2. $succ^{n+1}(m) = \bigcup_{m' \in succ^n(m)} succ(m')$.

Then $Inst^{\mathcal{T}} = \{succ^n(m_0) \,|\, n \in \mathbb{N}\}$ is the set of instants over $\mathcal{T}$ (we omit the superscript when the rooted DBT structure is clear from the context). We use $\mathtt{t}, \mathtt{t}_1, \mathtt{t}_2, \ldots$ to denote elements of $Inst^{\mathcal{T}}$.

According to Definition 4.2.3, each clock tick transitions every moment in an instant to the next unique instant.[4] When $m \in \mathtt{t}$ we say that *moment* m *occurs at instant* $\mathtt{t}$ and when $m \in h \cap \mathtt{t}$ we say that *history* h *crosses instant* $\mathtt{t}$ *at moment* m. Let $\langle Mom, m_0, < \rangle$ be a rooted DBT structure. The fact that $<$ is discrete and rooted in m_0 ensures that:

1. $Inst$ is a partition of Mom. Hence, every $m \in Mom$ occurs at one and only one instant, denoted with $\mathtt{t}_m$.

2. Every $h \in Hist$ crosses each instant $\mathtt{t}$ at exactly one moment, denoted with $m_{(\mathtt{t},h)}$. In what follows, we write $\mathtt{t}/h$ for $m_{(\mathtt{t},h)}/h$.

The above notation together with the notation introduced in Table 4.1 will be repeatedly used in Sections 4.3 and 4.4.

Turning to agency, we make the following two key assumptions about the individual actions that are performed at a moment: First, the action types in Atm, $Acts$, and $Ag\text{-}Acts$ represent *one-step actions*. So, in the spirit of PDL and CL, performing an action at a moment transitions to a set of *next* moments representing the different possible outcomes of the action.[5] Second, every transition from a moment to one of its successors is brought about by a unique complete group action. Accordingly, we label every index m/h with the complete group action that brings about the transition from m to its successor on h (i.e., the moment $succ_h(m)$). If index m/h is labeled with $\alpha \in Ag\text{-}Acts$, then $\alpha(i)$ represents *the action type that agent* $i \in Ag$ *instantiates at* m/h. Hence, every agent i instantiates one, and only one, type of action at every index m/h.

The final component of our semantics is a *deviant-choice function* that labels some of the individual actions that the agents can perform at a moment m as *deviant*. An action is deviant if it is not among the default actions that the agents can perform according to some underlying choice rules.

4.2.4. Definition (ALD_n frame). An ALD_n frame is a tuple $\langle \mathcal{T}, \mathbf{act}, \mathbf{dev} \rangle$ where $\mathcal{T}$ is a rooted DBT structure, $\mathbf{act} : Ind \to Ag\text{-}Acts$ assigns to every index a

[4] This is a convenient simplification, and is not essential for what follows. The crucial assumption is that, for every $m \in Mom$, there are alternative moments occurring at the same time as m.

[5] We think of the assumption that the temporal ordering is discrete as a by-product of this view of actions, rather than as an assumption about the structure of time in itself.

complete group action, and $\mathbf{dev} : Mom \to 2^{Acts}$ assigns to every moment a set of individual actions. For any $m \in Mom$ and $i \in Ag$, let

$$Acts_i^m = \bigcup_{h \in H_m} \mathbf{act}(m/h)(i)$$

be the set of individual actions *available to agent i at m* and

$$Acts^m = \bigcup_{i \in Ag} Acts_i^m$$

be the set of individual actions *executable at m*. The functions **act** and **dev** satisfy the following conditions: for all $m \in Mom$, $h, h' \in Hist$, and $i \in Ag$,

1. *No Choice Between Undivided Histories*: if h and h' are undivided at m, then $\mathbf{act}(m/h) = \mathbf{act}(m/h')$.
2. *Independence of Agents*: for all $\alpha \in Ag\text{-}Acts$, if $\alpha(j) \in Acts^m$ for all $j \in Ag$, then there is $h \in H_m$ such that $\mathbf{act}(m/h) = \alpha$.
3. *Executability of Deviant Actions*: $\mathbf{dev}(m) \subseteq Acts^m$.
4. *Availability of Default Actions*: there is $a_i \in Acts_i^m$ such that $a_i \notin \mathbf{dev}(m)$.
5. *Determinism of Choice Rules*: if there is an $a_i \in Acts_i \cap \mathbf{dev}(m)$, then $Acts_i^m \setminus \mathbf{dev}(m)$ is a singleton.

When $|Acts_i^m| = 1$, we say that agent i has a *vacuous choice* at m.

As we noted in Chapter 3.2.1, the set $Acts_i^m$ of actions available to agent i at moment m induces a partition on H_m: for every $h \in H_m$, the set

$$Acts_i^m(h) = \{h' \in H_m \mid \mathbf{act}(m/h)(i) = \mathbf{act}(m/h')(i)\}$$

is the cell in the partition containing h. The set $Acts_i^m(h)$ is the action token performed by i at m/h familiar in STIT semantics [see Chapter 2.1.1] that has been tagged with its assigned type. Note that every such action token is assigned a unique type and different tokens are assigned different types.[6]

Conditions 1 and 2 from Definition 4.2.4 are standard requirements in STIT semantics [see Chapter 2.1.1]: The condition of no choice between undivided histories ensures that no individual action executable at a moment can separate histories that are undivided at that moment. The condition of independence of agents ensures that every combination of individual actions executable at a moment (one for each agent) can itself be executed at that moment.

[6]This is a common idea and can be found in, e.g., Horty and Pacuit [2017]. It is also at the basis of the proof, presented by Broersen et al. [2006b], that CL [Pauly, 2002] can be embedded in STIT. See Chapter 2.3.1 for more details.

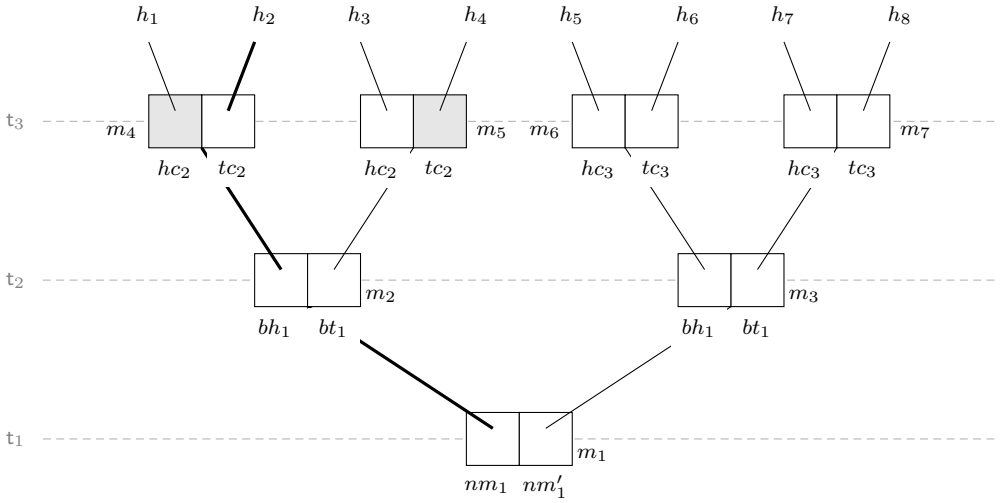


Figure 4.2: An ALD_n frame representing Example 4.1.1

The conditions on the function **dev** (conditions 3, 4 and 5 from Definition 4.2.4) require some more explanation. According to condition 3, only individual actions executable at a moment can be deviant at that moment. The idea is that individual actions that cannot be performed at a moment are immaterial for the default choice behavior of the agents at that moment. According to condition 4, every agent can perform at least one non-deviant action at every moment. Given the condition of independence of agents, this means that, at every moment, there is some history on which no agent performs a deviant action. So, according to the choice rules underlying an ALD_n frame, something will always happen.[7]

An ALD_n frame representing Example 4.1.1 is pictured in Figure 4.2. In the figure, David is agent 1, Max is agent 2, and Maxine is agent 3. David's individual action types are nm_1 (nominate Max), nm'_1 (nominate Maxine), bt_1 (bet tails), and bh_1 (bet heads); Max's individual action types are tc_2 (flip the T-coin) and hc_2 (flip the H-coin); and Maxine's individual action types are tc_3 (flip the T-coin) and hc_3 (flip the H-coin). The action types for agents with vacuous choices at a moment are omitted. The dashed lines represent instants and the gray cells represent the deviant actions (recall that Max's choice rule is that he flips the coin that guarantees that David loses).

We now define a model based on an ALD_n frame and truth of formulas from $\mathcal{L}_{\mathsf{ALD}_n}$ at an index and consider a potential axiomatization.

[7]This raises an immediate question: what if a moment has been reached by some agents performing deviant actions? We discuss this issue in Section 4.4.

(CPL)	Classical propositional tautologies	(MP)	From φ and $\varphi \to \psi$, infer ψ
($\mathsf{S5}_\Box$)	The axiom schemas of S5 for $\Box$	($\mathsf{RN}_\Box$)	From φ, infer $\Box\varphi$
(KD_X)	The axiom schemas of KD for X	(RN_X)	From φ, infer $\mathsf{X}\varphi$
(K_Y)	The axiom schemas of K for Y	(RN_Y)	From φ, infer $\mathsf{Y}\varphi$
	(I) Axioms for X and Y:		
(F_X)	$\hat{\mathsf{X}}\varphi \to \mathsf{X}\varphi$	(F_Y)	$\hat{\mathsf{Y}}\varphi \to \mathsf{Y}\varphi$
(C_{XY})	$\varphi \to \mathsf{X}\hat{\mathsf{Y}}\varphi$	(C_{YX})	$\varphi \to \mathsf{Y}\hat{\mathsf{X}}\varphi$
	(II) Axioms for *do*:		
(UH_{do})	$(do(\alpha) \wedge \mathsf{X}\Diamond\varphi) \to \Diamond(do(\alpha) \wedge \mathsf{X}\varphi)$	(Act)	$\bigvee_{a_i \in Acts_i} do(a_i)$
(IA_{do})	$(\Diamond do(a_1) \wedge \cdots \wedge \Diamond do(a_n)) \to \Diamond do(\alpha)$	(Sin)	$do(a_i) \to \neg do(b_i)$
	for $\alpha(1) = a_1, \ldots, \alpha(n) = a_n$		*for* $a_i \neq b_i$
	(III) Axioms for *dev*:		
(Ax1)	$\bigvee_{a_i \in Acts_i}(\Diamond do(a_i) \wedge \neg dev(a_i))$	(Ax3)	$dev(a_i) \to \Box dev(a_i)$
(Ax2)	$(\Diamond do(a_i) \wedge \Diamond do(b_i) \wedge \neg dev(a_i)) \to dev(b_i)$, *for* $a_i \neq b_i$	(Ax4)	$dev(a_i) \to \Diamond do(a_i)$

Table 4.2: A potential axiomatization of ALD_n

4.2.5. Definition (ALD_n model). Let *Prop* be defined as above. An ALD_n model is a tuple $\langle \mathcal{F}, \pi \rangle$, where $\mathcal{F}$ is an ALD_n frame and $\pi : Prop \to 2^{Ind}$ is a valuation function.

4.2.6. Definition (Truth for $\mathcal{L}_{\mathsf{ALD}_n}$). Suppose $\mathcal{M}$, is an ALD_n model. Truth of a formula $\varphi \in \mathcal{L}_{\mathsf{ALD}_n}$ at an index m/h in $\mathcal{M}$, denoted $\mathcal{M}, m/h \models \varphi$, is defined recursively. Truth of atomic propositions and the Boolean connectives is defined as usual. The remaining clauses are as follows:

$$
\begin{array}{lll}
\mathcal{M}, m/h \models do(a_i) & \text{iff} & \mathbf{act}(m/h)(i) = a_i \\
\mathcal{M}, m/h \models dev(a_i) & \text{iff} & a_i \in \mathbf{dev}(m) \\
\mathcal{M}, m/h \models \mathsf{X}\varphi & \text{iff} & \mathcal{M}, succ_h(m)/h \models \varphi \\
\mathcal{M}, m/h \models \mathsf{Y}\varphi & \text{iff} & m = m_0 \text{ or } \mathcal{M}, pred(m)/h \models \varphi \\
\mathcal{M}, m/h \models \Box\varphi & \text{iff} & \text{for all } h' \in H_m,\ \mathcal{M}, m/h' \models \varphi
\end{array}
$$

Table 4.2 displays a potential axiomatization of the logic ALD_n. The axioms and rules in the table extend the fragment of the axiom system ALO_n [cf. Table 3.2] without $\triangleright$ and merge a standard axiomatization for $\Box$, X, and Y with bridge principles connecting these modalities to the action description formulas using

do and *dev*. As noted in Chapter 3.2.2, the axioms for *do* are a reformulation, in $\mathcal{L}_{\mathsf{ALD}_n}$ (resp. $\mathcal{L}_{\mathsf{ALO}_n}$), of the main axioms of the logic $\mathcal{DLA}$ from Herzig and Lorini [2010]. The axioms in the last group express the fact that the function **dev** is moment-relative (axiom Ax1) and satisfies the conditions of executability of deviant actions (axiom Ax2), availability of non-deviant actions (axiom Ax3), and determinism of choice rules (axiom Ax4). It is not difficult to prove that the axioms and rules in Table 4.2 are valid and truth preserving in every ALD_n frame:

4.2.7. Theorem. *The axiom system defined in Table 4.2 is sound with respect to the class of all* ALD_n *frames.*

The proof of the following conjecture is the subject of ongoing research.

4.2.8. Conjecture. *The axiom system* ALD_n *is complete with respect to the class of all* ALD_n *frames.*

We anticipate that the proof of Conjecture 4.2.8 follows the same pattern as the completeness proof for the logic ALO_n presented in Appendix A.1. Some extra steps are needed in order to take care of the yesterday operator Y in the unraveling procedure presented in Appendix A.1.2.

4.2.3 Comparisons: strategic and epistemic STIT

To conclude this section, some brief comments about related extensions of STIT semantics are in order. The first extension that we discuss is strategic STIT [Belnap et al., 2001, Chapter 13; Horty, 2001, Chapter 7; Broersen and Herzig, 2015]. Labeling some actions as deviant at a moment can be viewed as a generalization of a strategy used in strategic STIT: Given a **dev** function and any agent i, we can define a function $s_i : Mom \to 2^{Acts_i}$ as follows: for all $m \in Mom$,

$$s_i(m) = \{a_i \in Acts_i^m \mid a_i \notin \mathbf{dev}(m)\}$$

Thus defined, s_i is a partial strategy for agent i that assigns to each moment m the non-deviant actions available to i at m. It is a *partial* strategy because some moment may be assigned all actions available to agent i at that moment. A similar generalization of strategic STIT can be found in Lorini and Sartor [2016], where the authors supplement STIT with a set of *rational choices* for every agent at every moment. But, as we mentioned in Section 4.1, choice rules may be grounded on preferences or habits that are, on the face of it, irrational. So, non-deviant choices may not coincide with rational choices. The approach that comes closest to our understanding of the **dev** function is Müller's [2005] idea of using strategic STIT to "affix 'defaults' to future choices" [*ibid.*, p. 199]. The key difference between Müller's proposal (and, more generally, strategic STIT) and our own is the role that "defaults" (or strategies) play in the semantics: "defaults"

are introduced here to contribute to the analysis of choice-driven counterfactuals rather than provide a semantics for strategic STIT operators.

A second extension of STIT adds epistemic operators [see, e.g., Broersen and Ramírez Abarca, 2018; Herzig and Troquard, 2006; Lorini et al., 2014; Horty and Pacuit, 2017]. It is important to not confuse an epistemic indistinguishability relation (an equivalence relation on indices) with instants. Our interpretation of instants is that they represent "alternative presents," and *not* uncertainty of the agents (or even the modeler). In fact, as it will become clear in Section 4.4, ALD_n frames are essentially extensive form games *with perfect information* (and simultaneous moves). This is important because instants will appear in the semantics of counterfactuals. In this chapter, we are interested in the truth conditions of choice-driven counterfactuals, and not what these counterfactuals may express about the cognitive procedure, knowledge, and beliefs that the agents' use to evaluate them. Such procedure will be the subject of Chapter 5.

4.3 Adding counterfactuals

In this section, we extend $\mathcal{L}_{\mathsf{ALD}_n}$ with formulas of the form $\varphi \boxrightarrow \psi$ with the interpretation "if φ were true, then ψ would be true." Let $\mathcal{L}^{\boxrightarrow}_{\mathsf{ALD}_n}$ be the full language. We aim at providing a semantics for $\mathcal{L}^{\boxrightarrow}_{\mathsf{ALD}_n}$ based on ALD_n frames. Our starting point is the well-known possible world semantics for counterfactuals due to Stalnaker [1968] and Lewis [1973a]. According to this approach, a counterfactual $A \boxrightarrow C$ is true at a world w just in case either

(i) there is no A-world accessible from w (the vacuous case), or

(ii) some $A \wedge C$-world is *more similar* to w than any $A \wedge \neg C$-world.[8]

The fundamental notion is a *relative similarity relation between possible worlds*, which Lewis [1973a] takes to be a *weak ordering* (a transitive relation in which ties are permitted but any two worlds are comparable) satisfying the *centering condition* (any world is more similar to itself than any other world).

The first question that arises when trying to adapt the mainstream semantics for counterfactuals to our semantics is: What should take the place of possible worlds as arguments of the relative similarity relation? In the Lewis-Stalnaker semantics, possible worlds are treated as unanalyzed entities. By contrast, in our framework formulas are interpreted at a moment on a history, which represents everything that happened in the past and everything that will happen in the future. From a logician's perspective, since Lewis defines relative similarity as a three-place relation on possible worlds and since indices (i.e., moment-history pairs) are the analogue of possible worlds in an ALD_n frame, relative similarity should be defined as a three-place relation over indices. However, when scholars

[8] Of course, for any sentence A, an A-world is a world satisfying A.

in the Lewisian tradition try to put flesh on the bones of Lewis's abstract relative similarity relation, they typically think of possible worlds as evolving over time (as *histories*) and not as momentary states (as moment-history pairs).[9] This squares, too, with the analysis of Example 4.1.1 we carried out in Section 4.1. In order to determine the truth value of

($C1$) If David had bet tails, then he would still have lost.

we consider *histories* that differed minimally from the actual one where it is true, at the time of utterance, that David bet tails and check whether, at that time, it is true that David loses. From this perspective, it makes sense to introduce a *relative similarity relation between histories* (rather than indices). We will see below that both perspectives can be accommodated.

Taking the more philosophical stance and following the intuitive analysis of Example 4.1.1, let us supplement ALD_n frames with a *relative similarity function*

$$\preceq\colon Hist \to 2^{Hist \times Hist}$$

that assigns to every history h a relative similarity relation $\preceq_h$, where, for all h, h_1, h_2,

$$h_1 \preceq_h h_2$$

means "h_1 is at least as similar to h as h_2." Let a *similarity* ALD_n *frame* be a tuple $\langle \mathcal{T}, \mathbf{act}, \mathbf{dev}, \preceq \rangle$ such that $\langle \mathcal{T}, \mathbf{act}, \mathbf{dev} \rangle$ is an ALD_n frame and $\preceq$ a relative similarity function. A *similarity* ALD_n *model* is a tuple $\langle \mathcal{T}, \mathbf{act}, \mathbf{dev}, \preceq, \pi \rangle$ where $\langle \mathcal{T}, \mathbf{act}, \mathbf{dev}, \preceq \rangle$ is a similarity ALD_n frame and π is a valuation function. Recall that, for any moment m, t_m is the instant to which m belongs (the time of m). When a formula is evaluated at m/h, we call t_m *the time of evaluation*. The following definition is the analogue of the Lewisian semantics for counterfactuals given above: a counterfactual is true at an index m/h just in case the consequent is true, at the time of evaluation t_m, on all histories that differ minimally from h where the antecedent is true at t_m (if there are any histories on which the antecedent is true at t_m).

4.3.1. Definition (Semantics for $\varphi \boxrightarrow \psi$). Where m/h is any index from a similarity ALD_n model $\mathcal{M}$ and $\varphi, \psi \in \mathcal{L}^{\boxrightarrow}_{\mathsf{ALD}_n}$,

$\mathcal{M}, m/h \models \varphi \boxrightarrow \psi$ iff either

(i) there is no $h_1 \in Hist$ such that $\mathcal{M}, \mathsf{t}_m/h_1 \models \varphi$, or

(ii) there is $h_1 \in Hist$ such that $\mathcal{M}, \mathsf{t}_m/h_1 \models \varphi \wedge \psi$ and, for all $h_2 \in Hist$ such that $\mathcal{M}, \mathsf{t}_m/h_2 \models \varphi \wedge \neg\psi$, $h_2 \not\preceq_h h_1$

[9] See, for instance, Lewis [1979] and Bennett [2003, Chapters 12-13].

A few definitions will clarify the connection between Definition 4.3.1 and the Lewis-Stanlanker semantics for counterfactuals. For any index m/h in a similarity ALD_n model $\langle \mathcal{T}, \mathbf{act}, \mathbf{dev}, \preceq, \pi \rangle$, let

$$\mathsf{t}(m/h) = \{m'/h' \in Ind \mid \mathsf{t}_m = \mathsf{t}_{m'}\}$$

be the set of indices *accessible* from m/h. So, an index m'/h' is accessible from m/h if it occurs at the same time as m/h. Next, for any index m/h, define $\preceq_{m/h} \subseteq Ind \times Ind$ by setting, for all $m_1/h_1, m_2/h_2 \in Ind$:

$$m_1/h_1 \preceq_{m/h} m_2/h_2 \text{ iff } m_1/h_1 \in \mathsf{t}(m/h) \text{ and } h_1 \preceq_h h_2.$$

That is, m_1/h_1 is at least as similar to m/h as m_2/h_2 just in case m_1/h_1 is accessible from m/h and h_1 is at least as similar to h as h_2. The evaluation rule for $\Box\!\!\rightarrow$ in Definition 4.3.1 can then be rewritten as:

$\mathcal{M}, m/h \models \varphi \Box\!\!\rightarrow \psi$ iff either

(i) there is no $m_1/h_1 \in \mathsf{t}(m/h)$ such that $\mathcal{M}, m_1/h_1 \models \varphi$, or

(ii) there is $m_1/h_1 \in \mathsf{t}(m/h)$ such that $\mathcal{M}, m_1/h_1 \models \varphi \wedge \psi$ and, for all $m_2/h_2 \in \mathsf{t}(m/h)$ such that $\mathcal{M}, m_2/h_2 \models \varphi \wedge \neg\psi$, $m_2/h_2 \not\preceq_{m/h} m_1/h_1$.

This is the standard evaluation rule for counterfactuals replacing possible worlds with indices. Rewriting Definition 4.3.1 in this way reveals two key assumptions underlying our semantics for counterfactuals.

The first assumption is that the truth values of φ and ψ at indices not occurring at the time of evaluation does not affect the truth-value of $\varphi \Box\!\!\rightarrow \psi$. This reflects the idea that, when we reason from a counterfactual supposition, we reason about what would happen if the supposed proposition were true *now* [cf. Thomason and Gupta, 1981, p. 68].[10] The second, more important, assumption is that the time of evaluation does not affect the relation of relative similarity between histories: if h_1 is at least as similar to h as h_2, then this is true *no matter what time it is*. This is a substantial assumption. Contrast it with what Thomason and Gupta [1981, pp. 68-69] call the condition of *past predominance*:

> (2.3) In determining how close m_1/h_1 is to m_2/h_2 [where m_1 and m_2 occur at the same time], past closeness predominates on future closeness; that is, the portions of h_1 and h_2 not after m_1 and m_2 predominate over the rest of h_1 and h_2.
>
> This informal principle is to be intended as strongly as possible: if h_3 up to m_3 is even a little closer to h_1 up to m_1 than is h_2 up to m_2,

[10] Another approach would be to tag each atomic proposition with the time of evaluation [see Shoham, 1989]. E.g., p_{t} means p is true at time t.

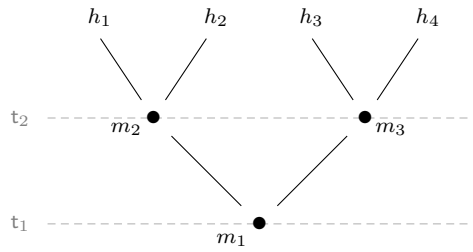


Figure 4.3: Past predominance, an illustration

> then m_3/h_3 is closer to m_1/h_1 than m_2/h_2 is, even if h_2 after m_2 is much closer to h_1 after m_1, than h_3 after m_3. Any gain with respect to the past counts more than even the largest gain with respect to the future. [Notation adapted.]

Consider the rooted DBT structure in Figure 4.3. The condition of past predominance implies that t_2/h_2 is more similar to t_2/h_1 than t_2/h_3, even if t_1/h_2 and t_1/h_3 are equally similar to t_1/h_1. This is excluded by our proposal, according to which, if t_2/h_2 is more similar to t_2/h_1 than t_2/h_3, then t_1/h_2 must be more similar to t_1/h_1 than t_1/h_3. The acceptance or rejection of the past predominance condition influences the logic of counterfactuals. We come back to this issue in Section 4.3.2.

4.3.1 Similarity defined

In this Section, we say more about the properties that our relative similarity relation $\preceq_h$ should satisfy.[11] We gradually introduce two candidate definitions of relative similarity in ALD_n frames. The first is based on Lewis's [1979] criteria for determining similarity and gives rise to what we call *rewind models*. The second, based on well-known counterexamples to Lewis's criteria [Slote, 1978, p. 27, fn.33], incorporates the idea that a notion of (in)dependence is key to a semantics of counterfactuals, giving rise to what we call *independence models*.

We start with Lewis's [1979, p. 472] first criterion of similarity: "It is of the first importance to avoid big, widespread, diverse violations of law." Lewis has

[11] As Bennett [2003, p. 196] notes:

> Lewis's theory evidently needs to be based [...] on a similarity relation that is constrained somehow – it must say that $A \boxdot\!\!\rightarrow C$ is true just in case C is true at the A-worlds that are most like the actual world in *such and such respects*. The philosophical task is to work out *what* respects of similarity will enable the theory to square with our intuitions and usage.

in mind mainly causal or physical laws, but the notion of law in the above quote can also be understood in terms of choice rules. The suggestion is that a history h_1 is more similar to a history h than another history h_2 if fewer deviations from the agents' default choice behavior occur on h_1 than on h_2. For any history h, the *number of deviations on* h is defined as follows:

$$n\_dev(h) = \sum_{m \in h} |\{i \in Ag \,|\, \mathbf{act}(m/h)(i) \in \mathbf{dev}(m)\}|$$

For any history h, $n\_dev(h)$ counts, for each moment m on h, the number of agents performing a deviant action at m/h. Our first analysis of relative similarity is:

Analysis 1. For all histories h, h_1, and h_2, h_1 is more similar to h than h_2 iff $n\_dev(h_1) < n\_dev(h_2)$.

Our first observation in this section is that our definition of similarity requires additional constraints that go beyond Analysis 1. To see this, consider again Example 4.1.1 and its representation in Figure 4.2. Recall that the actual history is h_2: after nominating Max, David bets heads and Max flips the T-coin, so David loses. Let L be the proposition that David loses (so, L is true at instant t_3 on h_2, h_3, h_6, h_7). We argued in Section 4.1 that the counterfactual $(C1)$ is true at m_4/h_2. The counterfactual $(C1)$ is expressed by the following formula of $\mathcal{L}_{\mathsf{ALD}_n}$:

$(F1)$ $\mathsf{Y}(do(bt_1)) \,\Box\!\!\rightarrow L$ ("If David had bet tails, then he would still have lost").

It is not hard to see that Definition 4.3.1 and Analysis 1 would evaluate $(F1)$ as false. The histories on which $\mathsf{Y}(do(bt_1))$ is true at the time of evaluation $\mathsf{t}_{m_4} = \mathsf{t}_3$ are h_3, h_4, h_7, and h_8. Among these histories, the ones with the fewest number of deviations are h_3, h_7, and h_8 (in fact, no deviant action is performed on these histories). But $\neg L$ rather than L is true on h_8 at t_3. So, if we compare histories only in terms of the number of deviations as in Analysis 1, then $(F1)$ turns out to be false at m_4/h_2. The problem with Analysis 1 is that it ignores the fact that a "small miracle" [Lewis, 1979] (or a "surgical intervention" [Pearl, 2000]) at m_4/h_2 suffices to reach h_3 from h_2, while a substantial change in the past is needed to reach h_7 and h_8. This suggests that *the greater past overlap between* h_1 *and* h_2 *is more important than the fewer number of deviations on* h_3.

Given the condition of past-linearity, the *past overlap between two histories* h_1 *and* h_2 is their intersection:[12]

$$past\_ov(h_1, h_2) = h_1 \cap h_2$$

This leads to a straightforward modification of Analysis 1:

[12]The condition of past-linearity ensures that $h_1 \cap h_2$ is an initial segment of both h_1 and h_2. This is why it makes sense to call it their *past* overlap.

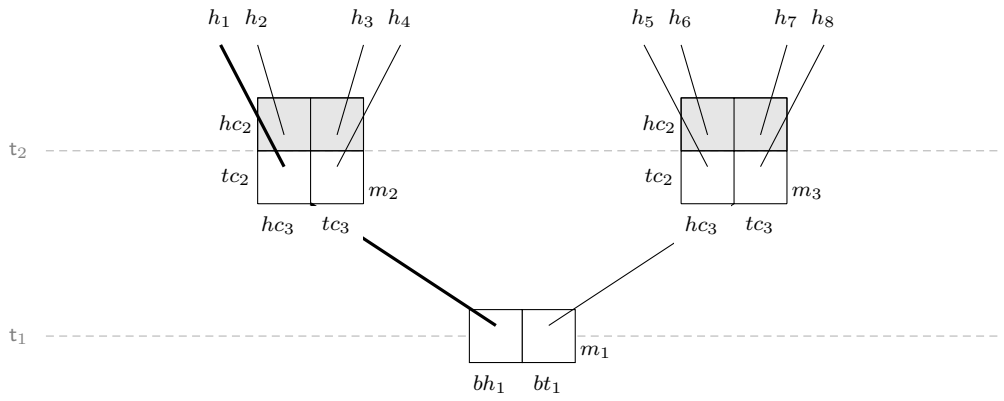


Figure 4.4: An ALD_n frame representing Example 4.3.2

Analysis 2. For all histories h, h_1, and h_2, h_1 is more similar to h than h_2 iff

either $past\_ov(h, h_1) \supset past\_ov(h, h_2)$,

or $past\_ov(h, h_1) = past\_ov(h, h_2)$ and $n\_dev(h_1) < n\_dev(h_2)$.[13]

Analysis 2 delivers the correct evaluation of $(F1)$ at m_4/h_2: Histories h_3 and h_4 are more similar to h_2 than h_7 and h_8, because their past overlap with h_2 is greater. In turn, history h_3 is more similar to h_2 than h_4 because there are fewer deviations on h_2 than on h_4. Since David loses at t_3 on history h_3, $(F1)$ is true at m_4/h_2. However, there are still problems with Analysis 2, as illustrated by the following example:

4.3.2. Example. Everything is as in Example 4.1.1 except that David does not initially select Max or Maxine. Instead, both Max and Maxine flip a coin after David bets. David wins only if both Max's and Maxine's coins land on the side he bets. Suppose that after David bets heads, Max flips the T-coin (as prescribed by his choice-rule) and Maxine happens to flip the H-coin. So David loses.

An ALD_n frame representing Example 4.3.2 is depicted in Figure 4.4, where the labels and shadings are read as in Figure 4.2 and the proposition L that David loses is true at instant t_3 on all histories except for h_2 and h_8. The actual history is h_1 (the thick line). Consider the following counterfactual:

$(F2)$ $do(hc_2) \,\Box\!\!\rightarrow \neg L$ ("If Max had flipped the H-coin, David would have won").

[13]The idea of using the notion of past overlap to define a relative similarity relation between histories in a branching time structure already appears in Xu [1997] and Placek and Müller [2007]. Unlike the present work, these papers do not consider any other criterion of similarity.

Intuitively, $(F2)$ is true at m_2/h_1. But Analysis 2 and Definition 4.3.1 do not vindicate this judgment. The histories on which Max flips the H-coin at $\mathsf{t}_{m_2} = \mathsf{t}_2$ are h_2, h_3, h_6, and h_7. Histories h_2 and h_3 have a greater past overlap with h_1 than h_6 and h_7, so the latter two histories can be discarded. In turn, since the number of deviations on h_2 is the same as the number of deviations on h_3, h_2 and h_3 are equally similar to h_1. Yet, L rather than $\neg L$ is true on h_3 at t_2. Given Definition 4.3.1, it follows that David *might* win – a weaker conclusion than the desired one. The problem is that, even though h_2 and h_3 have the same past overlap with h_1 as well as the same number of deviations, more agents need to change their actions to reach h_3 than h_2 (in this sense the change required to reach h_3 is not *minimal*).[14] This suggests that *the smaller change making h_2 branch off from h_1 is more important than the equal number of deviations on h_2 and h_3.*

Given two histories h_1 and h_2, say that h_1 and h_2 *divide at moment* m if m is the last moment they share, i.e., $m \in h_1 \cap h_2$ and $succ_{h_1}(m) \neq succ_{h_2}(m)$. When h_1 and h_2 divide at moment m, let the *number of agents separating h_1 and h_2* be defined as follows:

$$n\_sep(h_1, h_2) = |\{i \in Ag \,|\, \mathbf{act}(m/h_1)(i) \neq \mathbf{act}(m/h_2)(i)\}|$$

Then, $n\_sep(h_1, h_2)$ counts the number of agent that, by performing different actions on h_1 and h_2 at m, make h_1 and h_2 divide at moment m.[15] When h_1 and h_2 never divide (i.e., $h_1 = h_2$), let $n\_sep(h_1, h_2) = 0$. Putting everything together, we have our first definition of similarity.

4.3.3. Definition (Rewind similarity function). Let $\langle \mathcal{T}, \mathbf{act}, \mathbf{dev} \rangle$ be an ALD_n frame. Define

$$\prec^R : Hist \to 2^{Hist \times Hist}$$

by setting: for all $h, h_1, h_2 \in Hist$, $h_1 \prec^R_h h_2$ iff:

$$\begin{aligned}
&past_ov(h, h_1) \supset past_ov(h, h_2), \text{or} \\
&past_ov(h, h_1) = past_ov(h, h_2) \text{ and } n_sep(h, h_1) < n_sep(h, h_2), \text{or} \\
&past_ov(h, h_1) = past_ov(h, h_2) \text{ and } n_sep(h, h_1) = n_sep(h, h_2) \\
&\qquad\qquad\qquad\qquad \text{and } n_dev(h_1) < n_dev(h_2).
\end{aligned}$$

For every $h \in Hist$, define $\preceq^R_h$ as follows: for all $h_1, h_2 \in Hist$, $h_1 \preceq^R_h h_2$ iff either $h_1 \prec^R_h h_2$ or $(h_1 \not\prec^R_h h_2$ and $h_2 \not\prec^R_h h_1)$.

[14] The importance of fixing the actions of as many agents as possible when evaluating a counterfactual in a STIT model is already emphasized by Horty [2001, Chapter 4], who uses this criterion to define a selection function that picks, for every index m/h, agent i and action (token) K available to i at m, the most similar histories to h where i performs K. Since he is only interested in counterfactuals of form "if agent i performed (now) a different action, then φ would be true," Horty [2001] does not consider other criteria of similarity.

[15] Notice that, by the condition of past linearity, if two histories h_1 and h_2 divide at a moment, then they divide at a unique moment, so $n\_sep(h_1, h_2)$ is well defined.

We will call *rewind model* any similarity model $\langle \mathcal{T}, \mathbf{act}, \mathbf{dev}, \preceq^R, \pi \rangle$, where $\preceq^R$ is defined as in Definition 4.3.3.

Definition 4.3.3 encodes a substantial assumption about how we let a scenario unfold under the supposition that the antecedent of a counterfactual is true. To see this, let us go back to our initial Example 4.1.1, but suppose that the actual history is h_6 instead of h_2: After nominating Maxine, David bets heads and Maxine happens to flip the T-coin, so David loses. What if David had bet tails? Would he have won? There are two ways to answer this question.

1. *Rewind History*: When we suppose that David bet differently, we *rewind* the course of events to the moment when David bets (m_1), intervene on his choice, and then let the future unfold according to the agents' default choice behavior. Since there is no choice rule constraining Maxine's flip, we only conclude that David *might* win. This is the conclusion we reach by applying Definition 4.3.3, according to which h_3 and h_4 are equally similar to h_2. In fact, together with Definition 4.3.1, Definition 4.3.3 encodes the following Lewisian procedure:

   > [T]ake the counterfactual present, avoiding gratuitous difference from the actual present; graft it smoothly onto the actual past; let the situation evolve according to the actual laws; and see what happens. [Lewis, 1979, p. 463]

2. *Assume Independence*: When we suppose that David bet differently, we *only* intervene on his choice and leave all events that are *independent* of it as they actually are. Doing otherwise "would seem to be positing some strange causal influence" [Thomason and Gupta, 1981, p. 83]. Since there is no choice rule according to which Maxine's choice depends on David's bet, we conclude that, if David had bet differently, then he would have won.

To make the reasoning in item 2 precise, we need to describe relations of (in)depen-dence between the agents in ALD_n frames. Instead of introducing an additional parameter, we supplement Definition 4.3.3 with a further requirement on *unconstrained agents*. Recall that an agent i is unconstrained at a moment m when none of the actions available to her at m is deviant. Accordingly, we define the *set of agents unconstrained at moment* m as:

$$\underline{Ag}(m) = \{i \in Ag \mid Acts_i^m \cap \mathbf{dev}(m) = \varnothing\}$$

In terms of unconstrained agents, the idea underlying 2 is that, in reasoning from a counterfactual supposition, we do not change the actions of unconstrained agents.[16] Given an index m/h, define the *set of actions performed by unconstrained agents at* m/h as:

$$\underline{\mathbf{act}}(m/h) = \{\mathbf{act}(m/h)(i) \mid i \in \underline{Ag}(m)\}$$

[16] To account for the reasoning in 2, Thomason and Gupta [1981] impose constraints of "causal

For any histories h_1 and h_2, define:

$$n\_unc(h_1, h_2) = \sum_{\mathsf{t} \in Inst} |\underline{\mathbf{act}}(\mathsf{t}/h_1) \cap \underline{\mathbf{act}}(\mathsf{t}/h_2)|$$

Then, *n_unc* counts, for every instant t, the *number of agents unconstrained at* t *on both* h_1 *and* h_2 *that act in the same way on these histories*. Let us illustrate the previous definitions with Figure 4.2. Assume that the vacuous choices of agent $i \in \{1, 2, 3\}$ are all labeled with vc_i. We then have the following:

- $\underline{Ag}(m_k) = \{1, 2, 3\}$ for $k \in \{1, 2, 3, 6, 7\}$ and $\underline{Ag}(m_j) = \{1, 3\}$ for $j \in \{4, 5\}$;
- $\underline{\mathbf{act}}(\mathsf{t}_1/h_1) \cap \underline{\mathbf{act}}(\mathsf{t}_1/h_5) = \{vc_2, vc_3\}$,
  $\underline{\mathbf{act}}(\mathsf{t}_2/h_1) \cap \underline{\mathbf{act}}(\mathsf{t}_2/h_5) = \{bh_1, vc_2, vc_3\}$,
  $\underline{\mathbf{act}}(\mathsf{t}_3/h_1) \cap \underline{\mathbf{act}}(\mathsf{t}_3/h_5) = \{vc_1\}$,
  and so $n\_unc(h_1, h_5) = n\_unc(h_5, h_1) = 6$;
- $\underline{\mathbf{act}}(\mathsf{t}_1/h_5) \cap \underline{\mathbf{act}}(\mathsf{t}_1/h_7) = \{nm'_1, vc_2, vc_3\}$,
  $\underline{\mathbf{act}}(\mathsf{t}_2/h_5) \cap \underline{\mathbf{act}}(\mathsf{t}_2/h_7) = \{vc_2, vc_3\}$,
  $\underline{\mathbf{act}}(\mathsf{t}_3/h_5) \cap \underline{\mathbf{act}}(\mathsf{t}_3/h_7) = \{vc_1, vc_2, hc_3\}$,
  and so $n\_unc(h_5, h_7) = n\_unc(h_7, h_5) = 8$.

Our second definition of similarity incorporates the assumption of independence discussed in item 2 above.

4.3.4. Definition (Independence similarity function). Let $\langle \mathcal{T}, \mathbf{act}, \mathbf{dev} \rangle$ be an ALD_n frame. Define

$$\prec^I : Hist \to 2^{Hist \times Hist}$$

by setting: for all $h, h_1, h_2 \in Hist$, $h_1 \prec^I_h h_2$ iff either one of the first two conditions in Definition 4.3.3 is satisfied or one of the following holds:

$$\begin{aligned}
past_ov(h, h_1) = past_ov(h, h_2) &\text{ and } n_sep(h, h_1) = n_sep(h, h_2) \\
&\text{and } n_unc(h, h_1) > n_unc(h, h_2), \text{ or} \\
past_ov(h, h_1) = past_ov(h, h_2) &\text{ and } n_sep(h, h_1) = n_sep(h, h_2) \\
&\text{and } n_unc(h, h_1) = n_unc(h, h_2) \\
&\text{and } n_dev(h_1) < n_dev(h_2).
\end{aligned}$$

coherence" on their branching time models. Yet, they acknowledge that this move adds a substantial layer of complexity to their theory. With a similar aim but in the context of branching space-time, Placek and Müller [2007] define "independence" as space-like separation. Yet, they acknowledge that this kind of independence is hardly realized in everyday situations like the betting scenarios of our examples. The possibility of distinguishing constrained and unconstrained agents provides us with a convenient way to get around these difficulties.

For every $h \in Hist$, define $\preceq^I_h$ as follows: for all $h_1, h_2 \in Hist$, $h_1 \preceq^I_h h_2$ iff either $h_1 \prec^I_h h_2$ or ($h_1 \not\prec^I_h h_2$ and $h_2 \not\prec^I_h h_1$).

We will call *independence model* any similarity model $\langle \mathcal{T}, \mathbf{act}, \mathbf{dev}, \preceq^I, \pi \rangle$, where $\preceq^I$ is defined as in Definition 4.3.4.

4.3.5. Remark. In the rest of the chapter, we will use $\prec$ as a variable ranging over $\{\prec^R, \prec^I\}$. Similarly, we will use $\preceq$ as a variable ranging over $\{\preceq^R, \preceq^I\}$.

Definition 4.3.4 delivers the correct analysis of Example 4.3.2: although h_2 and h_3 overlap the same initial segment of h_1, at m_2 both David and Maxine act in the same way on h_2 and h_1, while Maxine changes her behavior on h_3. Hence, h_2 is more similar to h_1 than h_3. Since $\neg L$ is true on h_2 at t_2, it follows that $(F2)$ is true at m_2/h_1.

4.3.2 Logical properties

The following are some immediate consequences of Definitions 4.3.3 and 4.3.4:

4.3.6. Proposition. *Suppose that $\langle \mathcal{T}, \mathbf{act}, \mathbf{dev}, \preceq, \pi \rangle$ is either a rewind model or an independence model. For any history h, the relative similarity relation $\preceq_h$ is a centered weak ordering. That is, $\preceq_h$ satisfies the following conditions: for any $h', h_1, h_2, h_3 \in Hist$,*

1. *Transitivity: if $h_1 \preceq_h h_2$ and $h_2 \preceq_h h_3$, then $h_1 \preceq_h h_3$.*
2. *Linearity: either $h_1 \preceq_h h_2$ or $h_2 \preceq_h h_1$.*
3. *Centering: if $h' \preceq_h h$, then $h' = h$.*

Recall that, for any index m/h from a similarity ALD_n model, the set of indices accessible from m/h is $\mathsf{t}(m/h) = \{m'/h' \in Ind \mid \mathsf{t}_m = \mathsf{t}_{m'}\}$.

4.3.7. Corollary. *Suppose that $\langle \mathcal{T}, \mathbf{act}, \mathbf{dev}, \preceq, \pi \rangle$ is either a rewind model or an independence model. For any index m/h, the relation $\preceq_{m/h} \subseteq Ind \times Ind$ defined by setting: for all $m_1/h_1, m_2/h_2 \in Ind$,*

$$m_1/h_1 \preceq_{m/h} m_2/h_2 \text{ iff } m_1/h_1 \in \mathsf{t}(m/h) \text{ and } h_1 \preceq_h h_2,$$

is a centered weak ordering satisfying the following: for all $m_1/h_1, m_2/h_2 \in Ind$,

Priority: if $m_1/h_1 \in \mathsf{t}(m/h)$ and $m_2/h_2 \notin \mathsf{t}(m/h)$, then $m_2/h_2 \not\preceq_{m/h} m_1/h_1$

In tandem with Corollary 4.3.7, Definition 4.3.1 shows that our semantics for counterfactuals matches Lewis' semantics with possible worlds replaced with indices. Proposition 4.3.8 is then a well-known consequence of Corollary 4.3.7.

4.3.8. Proposition. *The following axioms and rule are valid and truth preserving in any rewind model (resp. independence model):*[17]

$(\mathsf{K}_{\Box\!\!\rightarrow})$	$(\varphi \Box\!\!\rightarrow (\psi_1 \rightarrow \psi_2)) \rightarrow ((\varphi \Box\!\!\rightarrow \psi_1) \rightarrow (\varphi \Box\!\!\rightarrow \psi_2))$
(Suc)	$\varphi \Box\!\!\rightarrow \varphi$
(Inc)	$(\neg\varphi \Box\!\!\rightarrow \varphi) \rightarrow (\psi \Box\!\!\rightarrow \varphi)$
(Cen)	$\varphi \rightarrow (\psi \leftrightarrow (\varphi \Box\!\!\rightarrow \psi))$
(Cond)	$(\varphi_1 \wedge \varphi_2 \Box\!\!\rightarrow \psi) \rightarrow (\varphi_1 \Box\!\!\rightarrow (\varphi_2 \rightarrow \psi))$
(RMon)	$\neg(\varphi_1 \Box\!\!\rightarrow \neg\varphi_2) \wedge (\varphi_1 \Box\!\!\rightarrow \chi) \rightarrow (\varphi_1 \wedge \varphi_2 \Box\!\!\rightarrow \chi)$
$(\mathsf{RN}_{\Box\!\!\rightarrow})$	*From* ψ *infer* $\varphi \Box\!\!\rightarrow \psi$

More interestingly, the principles in the next proposition reflect the interaction between counterfactuals and temporal modalities. The proof is in Appendix B.1.

4.3.9. Proposition. *The following principles are valid in any rewind model (resp. independence model).*

$(\mathsf{Dis}_{\mathsf{X}})$	$\mathsf{X}(\varphi \Box\!\!\rightarrow \psi) \leftrightarrow (\mathsf{X}\varphi \Box\!\!\rightarrow \mathsf{X}\psi)$	$(\mathsf{Dis}_{\mathsf{Y}})$	$\mathsf{Y}(\varphi \Box\!\!\rightarrow \psi) \leftrightarrow (\mathsf{Y}\varphi \Box\!\!\rightarrow \mathsf{Y}\psi)$
(Cen1)	$\Diamond\varphi \rightarrow (\Diamond\psi \rightarrow \Box(\varphi \Box\!\!\rightarrow \Diamond\psi))$	(Cen2)	$\Diamond\varphi \rightarrow ((\varphi \Box\!\!\rightarrow \Box\psi) \rightarrow \Box\psi)$

4.3.10. Corollary. *The following are theorems of the axiom system defined by the axioms and rules in Table 4.2, the principles in Proposition 4.3.8, and principles* Cen1 *and* Cen2 *from Proposition 4.3.9:*

1. $\Diamond\varphi \rightarrow (\Box\psi \leftrightarrow \Box(\varphi \Box\!\!\rightarrow \Box\psi))$,
2. $\Diamond\varphi \rightarrow ((\varphi \Box\!\!\rightarrow \Box\psi) \rightarrow \Box(\varphi \Box\!\!\rightarrow \psi))$.

Proof:
Straightforward given Cen1, Cen2 and the fact that $\Box$ is an S5 modality. □

The validity of the distribution principles $\mathsf{Dis}_{\mathsf{X}}$ and $\mathsf{Dis}_{\mathsf{Y}}$ depends on the assumption that the time of evaluation does not affect the relation of relative similarity between histories. In fact, since the most similar histories to a history h *up to the present time* t are the same as the most similar histories to h *up to one instant after* t, the most similar histories to h on which $\mathsf{X}\varphi$ is true *at* t must be the same as the most similar histories to h on which φ is true *one instant after* t (similarly for $\mathsf{Y}\varphi$). Thomason's and Gupta's [1981] principle of past predominance makes it possible to find counterexamples to $\mathsf{Dis}_{\mathsf{X}}$ and $\mathsf{Dis}_{\mathsf{Y}}$. To see this, let us go back to Figure 4.3. Recall that, according to the condition of past

[17] Suc stands for "Success," Inc for "Inclusion" (as it says that all indices satisfying a counterfactual antecedent are accessible), Cen stands for "Centering," Cond for "Conditionalization," and RMon for "Rational Monotonicity."

predominance, t_2/h_2 is more similar to t_2/h_1 than t_2/h_3, while t_1/h_2 and t_1/h_3 are equally similar to t_1/h_1. Suppose that p is true only at t_2/h_2 and t_2/h_3 and that q is true only at t_2/h_2. Since q is true at the most similar index to t_2/h_1 at which p is true (i.e., t_2/h_2), $p \boxdot\!\!\rightarrow q$ is true at t_2/h_1, and so $\mathsf{X}(p \boxdot\!\!\rightarrow q)$ is true at t_1/h_1. On the other hand, since $\neg\mathsf{X}q$ is true at one of the most similar indices to t_1/h_1 at which $\mathsf{X}p$ is true (i.e., t_1/h_3), $\mathsf{X}p \boxdot\!\!\rightarrow \mathsf{X}q$ is false at t_1/h_1.

Thomason and Gupta [1981, pp. 70-71] rely on a variant of Example 4.1.1 to support the claim that Dis_X and Dis_Y should not come out as logical validities. In their version of the example, Max and David are the only agents, the game starts with David's bet (at t_2 in Figure 4.2) and ends after Max flips either the T-coin or the H-coin (after t_3 in Figure 4.2). As in Example 4.1.1, Max flips the coin that guarantees that David loses. In addition, the actual history is h_2: David bets heads and Max flips the T-coin. Now, let L' be the proposition "David loses at time t_3" (so, L' is true at all moments on histories h_2, h_3, h_6, and h_7). According to Thomason and Gupta [1981], the counterfactual

(A) $do(bt_1) \boxdot\!\!\rightarrow L'$ ("If David bets tails, he would lose at t_3")

is *intuitively true* at t_2/h_2, i.e., *at the beginning of the game* on the actual history. Hence, $\mathsf{Y}(do(bt_1) \rightarrow L')$ is true at t_3/h_2. On the other hand, the authors take the counterfactual

(B) $\mathsf{Y}do(bt_1) \boxdot\!\!\rightarrow \mathsf{Y}L'$ ("If David had bet tails, he would have lost at t_3")

to be *intuitively false* at t_3/h_2, i.e., *at the end of the game* on the actual history. If this is correct, then the implication $\mathsf{Y}(do(bt_1) \boxdot\!\!\rightarrow L') \rightarrow (\mathsf{Y}do(bt_1) \boxdot\!\!\rightarrow \mathsf{Y}L')$ is false at t_3/h_2, that is, the principle DIS_Y is *not* intuitively valid.

We disagree with Thomason's and Gupta's judgment about (B). Given Max's choice rule, at the end of the game it would be perfectly natural to explain to David: "Well, if you had bet tails, you would still have lost." We think that the problem stems from a confusion between the *time of evaluation* and the *time to which the antecedent of a counterfactual refers*. In discussing the present example, Thomason and Gupta seem to take it that, in reasoning from a counterfactual supposition, we hold fixed as many past facts as possible up to the time of evaluation (t_1 in the case of (A) and t_2 in the case of (B)). But, as most scholars think [cf. Bennett, 2003, Chapter 12], what we intuitively do is rather to hold fixed as many past facts as possible up to the time to which the antecedent refers (t_1 for both (A) and (B)).[18] It then makes sense that relative similarity between histories is not affected by the time of evaluation: what is important is just that the longer a history h' overlaps another history h, the more similar h' is to h.

[18] Observe that, if we kept fixed as many past facts as possible up to the time of evaluation, (B) would be false, no matter whether Max flips the T-coin by chance or because his default choice behavior is to make David lose. Yet, intuitively, we judge (B) false only in the former case (recall the reasoning underlining the "Rewind history" and "Assume independence" attitudes).

Turning to Cen1 and Cen2, the validity of these principles follows from the priority of the criterion of past overlap: if φ *can* be true at a *moment*, then supposing that φ is true does not require shifting to a different moment. (Compare the reasoning behind the validity of Cen: if φ *is* true at an *index*, then supposing that φ is true does not require moving to a different index.) Items 1 and 2 in Theorem 4.3.10 highlight an interesting interaction between counterfactuals and historical necessity. In particular, item 2, which we discuss below, can be viewed as a principle of "exportation" of $\Box$ from $\Box\!\!\rightarrow$.

The validities we have considered so far do not depend on whether we work with rewind models or with independence models. The next Proposition 4.3.11 involves a formula that distinguishes the two classes of models. The proof can be found in Appendix B.2.

4.3.11. Proposition. *The following principle is valid in any rewind model, but invalid in some independence models.*

$$(\mathsf{Exp}_{\Box}) \quad \Box\neg\varphi \rightarrow ((\varphi \,\Box\!\!\rightarrow \Box\psi) \rightarrow \Box(\varphi \,\Box\!\!\rightarrow \psi))$$

Using item 2 in Theorem 4.3.10 and $\mathsf{Exp}_{\Box}$ we can show that $(\varphi \,\Box\!\!\rightarrow \Box\psi) \rightarrow \Box(\varphi \,\Box\!\!\rightarrow \psi)$ is valid in the class of rewind models. The validity of this principle can be proved directly from Definition 4.3.3, which ensures that the most similar φ-histories[19] to histories passing through a moment pass through the same moments. Note that the converse implication is not valid: Suppose that we scheduled a lecture on Tuesday at 1pm and our default choice behavior is to follow the schedule. Then, "If I were not sick, I would be teaching" is settled true on Tuesday at 1pm, even though "If I were not sick, it would be settled that I would be teaching" may be false (e.g., because there is a possibility that our bike breaks down on the way to school).

To see why the addition of the criterion regarding unconstrained agents leads to the invalidity of $\mathsf{Exp}_{\Box}$, consider another example.

4.3.12. Example. Charles puts an apple, a banana, an orange, and a grapefruit in a basket and three pieces of paper with the choices *orange+grapefruit*, *orange+apple*, *grapefruit+banana* written on them in a jar. Bob picks the *grapefruit+orange*-paper, and receives the corresponding fruits. After that, Charles splits the remaining pieces of paper in half. Ann picks the *banana*-paper, and receives the corresponding fruit.

Example 4.3.12 can be modeled as illustrated in Figure 4.5. In the figure, Bob is agent 1 and his non-vacuous choices are og_1 (pick the *orange+grapefruit*-paper), oa_1 (pick the *orange+apple*-paper), and gb_1 (pick the *grapefruit+banana*-paper). Ann is agent 2 and her non-vacuous choices are a_2 (pick the *apple*-paper), b_2

[19] By "φ-history" we mean a history on which φ is true at the time of evaluation.

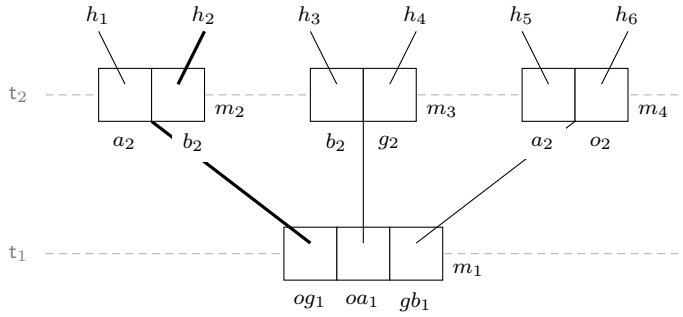


Figure 4.5: An ALD_n frame representing Example 4.3.12

(pick the *banana*-paper), g_2 (pick the *grapefruit*-paper), and o_2 (pick the *orange*-paper). The actual history (thick line) is h_2. In our terminology, both Bob and Ann are unconstrained agents – none of their actions is deviant with respect to any preference or strategy. At m_2, there are no citrus fruits in the basket. But what if there were? According to Definition 4.3.4, the most similar history to h_2 satisfying this condition is h_3, where Bob picks the *orange+apple*-paper and Ann the *banana*-paper – as she does at m_2/h_2. At $\mathtt{t}_2/h_3$ it is settled that Ann can pick a banana, so "If there were a citrus fruit in the basket, it would be settled that Ann could pick a banana" is true at m_2/h_2. But consider index m_2/h_1 where Ann picks the *apple*-paper instead of the *banana*-paper. What if there were a citrus fruit in the basket? Reasoning as before, the most similar history to h_1 satisfying this condition is h_5, where Bob picks the *grapefruit+banana*-paper and Ann the *apple*-paper. Since there is no banana in the basket at $\mathtt{t}_2/h_5$, "If Ann could pick a citrus fruit, she could pick a banana" is false at m_2/h_1, and so "It is settled that, if there were a citrus fruit left, Ann could pick a banana" is false at m_2/h_2.

Before proceeding, let us highlight a potential problem for our proposal emerging from Figure 4.5. We have seen that, according to Definition 4.3.4, h_3 is the most similar history to h_2 on which Bob does not choose the *orange+grapefruit*-paper. So, "If Bob had picked a different piece of paper, then Ann would pick the *banana*-paper" is true at m_2/h_2. But this is counterintuitive: if Bob had picked a different piece of paper, he might have picked the *grapefruit+banana*-paper, in which case Ann could not even pick the *banana*-paper! We view this as a modeling issue: choosing a *banana*-paper over a *apple*-paper is not the same type of choice as choosing a *banana*-paper over a *grapefruit*-paper, so the two choices should not be labeled the same way.[20] If we change the labeling, then the weaker (and unproblematic) "If Bob had picked a different piece of paper, then Ann might

[20]See the discussion of menu dependence in rational choice theory [Dietrich and List, 2016; Kalai et al., 2002; Sen, 1997].

pick the banana-paper" is true at m_2/h_2. This suggests the introduction of the next condition: for all $i \in Ag$ and $m, m' \in Mom$,

1. *Identity of Overlapping Menus*: if $Acts_i^m \cap Acts_i^{m'} \neq \varnothing$, then $Acts_i^m = Acts_i^{m'}$.

Interestingly, as proved in Appendix B.2, $\mathsf{Exp}_\Box$ remains invalid in the class of independence models satisfying the above condition 1. In fact, the countermodel presented there satisfies a stronger condition: for all $m, m' \in Mom$,

2. *Uniformity of Menus*: if $\mathsf{t}_m = \mathsf{t}_{m'}$, then $Acts^m = Acts^{m'}$.

While the condition of identity of overlapping menus is a desirable condition, the condition of uniformity of menus is not: as illustrated by Example 4.3.12, depending on what happens at a moment, different actions may become executable in the future. But is the condition of identity of overlapping menus enough to eliminate counterintuitive results? We leave a full exploration of this issue to future research.

4.4 Deviant choices and counterfactuals

All of the examples discussed in the previous sections involve counterfactuals evaluated at indices at which all of the agents choose actions prescribed by their choice rules. How should we evaluate choice-driven counterfactuals at moments arrived at by some agents performing deviant actions? Consider the following variant of our running example.

4.4.1. Example. Everything is as in Example 1, except that, besides the two biased coins, Max also can choose a fair coin – and he knows this. Max's choice rule is the same: choose the coin the guarantees that David will lose. Suppose that David nominates Max and bets heads but Max makes a mistake and flips the fair coin, which, lucky for David, lands heads.

Example 4.4.1 is depicted in Figure 4.6. Max's choice of flipping the fair coin is represented by the action type fc_2. The coin lands heads on histories h_2 and h_6 and lands tails on histories h_3 and h_7. At the index t_2/h_2, how should we evaluate the counterfactual $(C1)$ discussed in Section 4.1?

$(C1)$ If David had bet tails, then he would still have lost. $(\mathsf{Y}(do(bt_1)) \boxminus\!\!\rightarrow L)$

To evaluate $(C1)$, we need to determine which coin Max would flip if David had just bet tails. Since Max is constrained by his choice rule, both Definition 4.3.3 and Definition 4.3.4 deliver the same analysis: after rewinding and intervening on David's choice, the future unfolds "forgetting" that David's choice was deviant on the actual history. This implies that $(C1)$ is true at t_2/h_2 since Max would

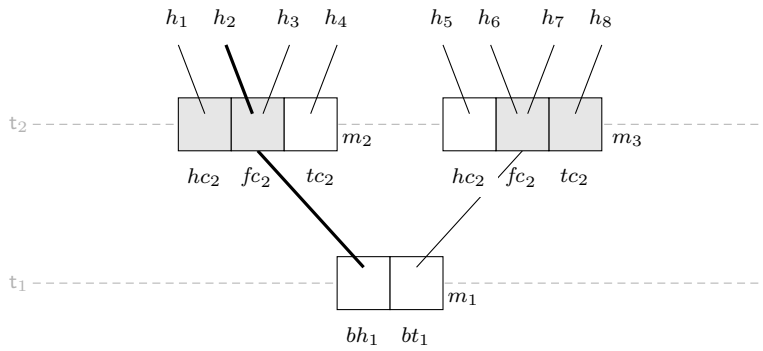


Figure 4.6: An ALD_n frame representing Example 4.4.1

flip the H-coin at m_3, as prescribed by his choice rule. It is not clear that this is the correct judgment about $(C1)$ given that Max mistakenly flipped the fair coin.

The main issue is that neither definition of similarity takes into account the fact that the counterfactual is evaluated on a history on which Max acted as an *unconstrained* agent. There are different ways to determine what coin Max will flip under the supposition that David bet differently:

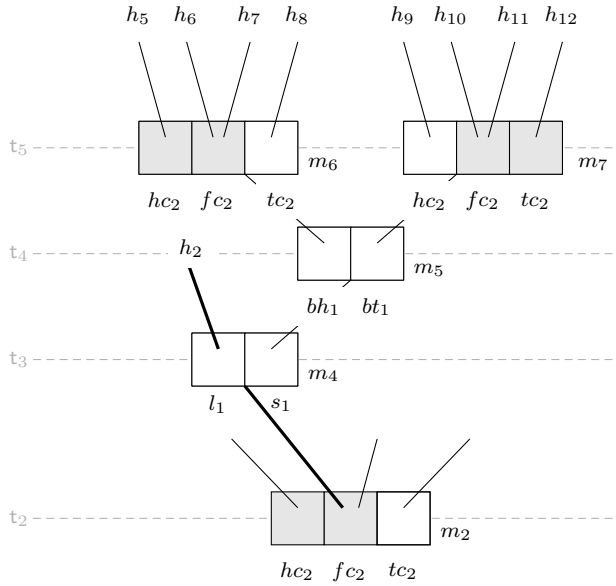
1. Forget that Max's actual choice was deviant and assume that he is constrained by his choice rule. (This is what Definition 4.3.3 and Definition 4.3.4 implicitly assume.)

2. Assume that Max would have made the same mistake and flip the fair coin.

3. Assume that Max would have made *a* mistake, but we cannot tell which one (e.g., he *might* flip the fair coin or the tails coin).

4. Assume that Max is no longer a constrained agent, so the only conclusion we can draw is that Max might flip *any* of the available coins.

Without further details about why Max acted deviantly, it is not clear which of the above options is best. Perhaps Max made a fleeting mistake and there is no further explanation, which would suggest that option 1 is the best. There might be a systematic problem with the coins (e.g., they are labeled incorrectly), which would suggest that option 2 is the best. Finally, options 3 and 4 are best if Max's deviant action is some type of signal that he is no longer being guided by his choice rule. Distinguishing these options is particularly important when evaluating "forward-looking" counterfactuals involving statements about what will happen in the future. Consider the following modification of Example 4.4.1.



The diagram extends history h_2 in Figure 4.6 with a choice for David at m_4 to leave (l_1) or stay (s_1), followed by another round of the game described in Example 4.4.1.

Figure 4.7: An ALD_n frame representing Example 4.4.2

4.4.2. Example. Everything is as in Example 4.4.1, except that David can choose to either leave or stay and play another round of the game after Max flipped his coin. As in Example 4.4.1, suppose that in the first game David bets heads and then Max makes a mistake by flipping the fair coin, which lands heads. After this game, David chooses to leave the game.

Example 4.4.2 is depicted in Figure 4.7. The actual history is h_2 (the thick line): David bets heads, then Max mistakenly flips the fair coin (which lands heads), and finally David decides to not play another round of the game. How should we evaluate the following counterfactual ($C2$) at $\mathtt{t}_3/h_2$?

($C2$) If David were to bet heads, he would win.

Let W stand for "David wins the second game," so W is true at $\mathtt{t}_5/h_5$, $\mathtt{t}_5/h_6$, $\mathtt{t}_5/h_{11}$, and $\mathtt{t}_5/h_{12}$. Then, ($C2$) is represented by the formula:

($F2$) $\mathsf{X}do(bh_1) \,\Box\!\!\rightarrow \mathsf{XX}W$.

Note that $\mathsf{X}do(bh_1)$ is false at $\mathtt{t}_3/h_2$.[21] According to either Definition 4.3.3 or Definition 4.3.4, ($F2$) is false at $\mathtt{t}_3/h_2$: the most similar history to h_2 on which

[21] Of course, the successor of m_4 on h_2 is not represented in Figure 4.7. It is assumed that the game is over at m_4/h_2 and so the next choice for David on h_2 does not involve betting heads.

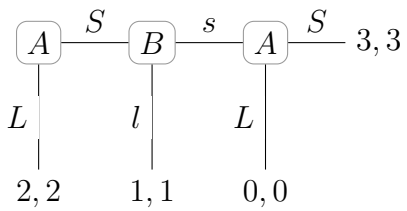
David bets heads during the second game is h_8, where $\mathsf{XX}W$ is false at t_3. This means that the following is also false at t_3/h_2:

($C3$) If David were to bet heads, then he *might* win.

The main reason why ($C2$) and ($C3$) are false at t_3/h_2 is because it is assumed that Max's choice at m_6 is not influenced in any way by his errant choice at m_2. However, if the circumstances that led to Max's deviant choice at m_2 are still in place at m_6, then, arguably, ($C3$) should be true at t_3/h_2. The upshot is that the evaluation of counterfactuals after one or more deviant choices often depends on making precise details about why the agents did not choose according to their choice rules. Drawing on some work in the foundations of game theory, we show how to extend our models to represent some of these details.

Counterfactuals, such as ($C2$) and ($C3$), play an important role in the analysis of strategic reasoning in game theory [Selten and Leopold, 1982; Bicchieri, 1988; Shin, 1992; Stalnaker, 1996; Skyrms, 1998; Zambrano, 2004; Bonanno, 2015]. A central question in this literature is: What do the players expect that their opponents will do if an unexpected point in the game tree is reached? One answer (forward induction) is that players rationalize past behavior and use it as a basis for forming beliefs about future moves [Battigalli, 1997; Battigalli and Siniscalchi, 2002; Stalnaker, 1998]. A second answer (backward induction) is that players ignore past behavior and reason only about their opponents' future moves [Aumann, 1995; Bonanno, 2014; Perea, 2014; Stalnaker, 1998]. These different answers roughly correspond to the four different options listed above that make precise why Max made a deviant choice. Forgetting that Max made a deviant choice and assuming he will be guided by his choice rule (option 1) is analogous to the assumptions underlying backward induction reasoning (the second answer). The other options can be viewed as different ways to rationalize Max's surprising choice, as in forward induction reasoning.

We leave to further work a more detailed discussion about strategic reasoning in STIT models, and focus on the question raised above about how to evaluate choice-driven counterfactuals after one or more deviant choices. This question is related to an issue that arises when developing epistemic characterizations of solutions concepts for *extensive form games* [Pacuit and Roy, 2015; Perea, 2012]. We illustrate this issue with the following example:



This is a game between two players Alice (A) and Bob (B). Player A moves first and can choose between leave (L) and stay (S). If A chooses L, then the

game ends with a payout of 2 to both players. If A chooses S, then player B chooses between l and s. If B chooses l, then the game ends with a payout of 1 to both players; and if B chooses s, then player A chooses a second time. If A chooses L at her second decision node, then the game ends with both players receiving a payout of 0, and if A chooses S at her second decision node, then the game ends with both players receiving a payout of 3. A *strategy for player i* in an extensive form game is a function that assigns an action to each decision node for player i. Importantly, a strategy s for player i specifies actions that player i would take at *all* decision nodes, including those that are not reachable given the actions assigned to i at earlier decision nodes. For example, choosing L at the first decision node and S at the second decision node is a strategy for player A.

Observe that ALD_n models are essentially extensive form games except that we allow more than one player to move at a given moment.[22] In our terminology, a strategy for player i is a choice rule, so strategies are represented in our models by a **dev** function.

The *backward induction solution* in the above game is for A to choose S at her first decision node, for B to choose s at his decision node, and finally for A to choose S at her second decision node, resulting in a payout of 3 to both players. Stalnaker [1998] used the above game to show, contrary to a famous result by Aumann [1995], that the backward induction strategy is not the only one consistent with the players commonly believing that everyone is *rational*. We can represent the key difference between Aumann and Stalnaker as a disagreement about the judgment of a counterfactual [cf. Halpern, 2001]: Fix a choice rule for Alice that says to choose L at her first decision node and S at her second decision node. Consider the following statements:

1. If Alice were to choose (at her second decision node), she would choose L

2. If Alice were to choose (at her second decision node), she would choose S

According to Aumann, statement 2 is true *no matter what Alice does at her first decision node*. That is, Alice's behavior at her first decision node has no influence over her choice at her second decision node. Stalnaker disagrees. He argues that if Alice (contrary to her given choice rule) chose S at the first decision node, then statement 2 is false:

> [S]ince she was irrational once, she will be irrational again, and so would choose [L]. (It would be enough if [Bob] concluded only that for all he then would know, she *might* act irrationally again.) [Stalnaker, 1998, p. 47]

[22] That is, our models are extensive form games with simultaneous moves and perfect observation [Osborne and Rubinstein, 1994]. Of course, a key difference is that we do not represent any utility information for the agents.

There is more to Stalnaker's argument as he justifies why Alice's choice rule given above is *rational*. Since here we are not interested in the rationality of choice rules, the key point is that Alice's choice of S at her first decision node is deviant. We can then rephrase in our terminology the different judgements of Aumann and Stalnaker as follows: Aumann assumes that past deviant behavior has no influence on choices at later moments. Stalnaker assumes that if an agent acted deviantly in the past, then she may act deviantly at a later moment. As explained above, both Definition 4.3.3 and Definition 4.3.4 incorporate Aumann's assumption. To incorporate Stalnaker's assumption, we must refine our models to keep track of which choices are influenced by earlier deviant choices.

The main idea is to refine our definition of similarity by taking into account that a deviant action at a moment might imply deviant actions at some later moments, possibly on alternative, hypothetical histories [cf. item 3, p. 101]. To do this, we introduce a relation $\rightsquigarrow$ between agent-moment pairs representing the influence that a deviant choice by an agent i at a moment m has on the choices of the various agents at such "alternative" moments. Let $(i, m) \rightsquigarrow (j, m')$ mean that, if i chooses deviantly at m, then j would choose deviantly at m'.

4.4.3. Definition (Influence). Let $\mathcal{F} = \langle \mathcal{T}, \mathbf{act}, \mathbf{dev} \rangle$ be an ALD_n frame and Ag be defined as above. An *influence relation* for Ag in $\mathcal{F}$ is a relation

$$\rightsquigarrow \subseteq (Ag \times Mom) \times (Ag \times Mom)$$

such that, for all $(i, m), (j, m') \in Ag \times Mom$, if $(i, m) \rightsquigarrow (j, m')$, then $m'' \leq m'$ for some $m'' \in \mathsf{t}_m$.

The influence relation captures how deviant choices spread on a history. The only constraint on the influence relation is that an agent's deviant choice at a moment m can only influence agents' choices at moments occurring either at t_m or at later instants. To incorporate this into our definition of similarity, let $n\_inf(h, h')$ be the number of agents choosing deviantly at a moment on h' that are influenced by a deviant choice of an agent at a moment on h. For histories h and h', $n\_inf(h, h')$ is defined as follows:

$$n\_inf(h, h') = \sum_{m \in h} |\{(j, m') \in Ag \times h' \mid \text{there is } i \in Ag \text{ s.t. } \mathbf{act}(m/h)(i) \in \mathbf{dev}(m), \\ (i, m) \rightsquigarrow (j, m') \text{ and } \mathbf{act}(m'/h')(j) \in \mathbf{dev}(m')\}|$$

Using n_inf, we can modify Definition 4.3.3 to account for the deviant choices of some agents.

4.4.4. Definition (Similarity given a deviant action). Let $\langle \mathcal{T}, \mathbf{act}, \mathbf{dev} \rangle$ be an ALD_n frame. Define

$$\prec^D : Hist \to 2^{Hist \times Hist}$$

by setting: for all $h, h_1, h_2 \in Hist$, $h_1 \prec^D_h h_2$ iff either one of the first two conditions in Definition 4.3.3 is satisfied or one of the following holds:

$$\begin{aligned} past\_ov(h,h_1) = past\_ov(h,h_2) &\text{ and } n\_sep(h,h_1) = n\_sep(h,h_2) \\ &\text{and } n\_inf(h,h_1) > n\_inf(h,h_2), \text{ or} \\ past\_ov(h,h_1) = past\_ov(h,h_2) &\text{ and } n\_sep(h,h_1) = n\_sep(h,h_2) \\ &\text{and } n\_inf(h,h_1) = n\_inf(h,h_2) \\ &\text{and } n\_dev(h_1) < n\_dev(h_2). \end{aligned}$$

For any $h \in Hist$, a relation $\preceq_h^D$ can be defined as in Definition 4.3.3.

4.4.5. Remark. Note that both Definition 4.3.4 and Definition 4.4.4 modify Definition 4.3.3 in similar ways. Indeed, Definition 4.4.4 is the same as Definition 4.3.4 with *n_unc* replaced with *n_inf*.

Consider the STIT model in Figure 4.6. Suppose that $(2, m_2) \rightsquigarrow (2, m_3)$. That is, if 2 chooses deviantly at m_2, then 2 would choose deviantly at m_3. Then, $n\_inf(h_2, h_5) = 0$ since 2 chooses deviantly at m_2/h_2, $(2, m_2) \rightsquigarrow (2, m_3)$, but 2 does not choose deviantly at m_3/h_5. Since 2 chooses deviantly at $m_2/h_6, m_2/h_7$ and m_2/h_8, we have $n\_inf(h_2, h_6) = n\_inf(h_2, h_7) = n\_inf(h_2, h_8) = 1$. Thus, according to Definition 4.4.4, h_6, h_7 and h_8 are equally similar relative to h_2 and all strictly more similar to h_2 than h_5 (i.e., $h_6 \prec_{h_2}^D h_5$, $h_7 \prec_{h_2}^D h_5$, $h_8 \prec_{h_2}^D h_5$). This means that $(C1)$ is false at m_2/h_2. Similarly, in Figure 4.7, if $(2, m_2) \rightsquigarrow (2, m_6)$, then $(C2)$ is false at m_4/h_2, but $(C3)$ is true at m_4/h_2.

Returning to the extensive form game given above, Stalnaker assumes that Alice's deviant choice at her first decision node influences her deviant choice at her second decision node. This explains the judgment that statement 2 is false. On the other hand, Aumann does not assume that a deviant choice at the first decision node influences what Alice chooses at her second decision node. This explains the judgment that statement 2 on page 104 is false.

4.5 Conclusion

In this chapter, we studied the semantics and logical properties of choice-driven counterfactuals in a STIT logic that improves the logic ALO_n from Chapter 3 by allowing us to describe past facts, properties of instants, and deviant actions. Following Lewis [1973a], we interpreted counterfactual statements using a relation of relative similarity on histories. We introduced three definitions of similarity motivated by different intuitions about how choice rules guide the agents' actions in counterfactual situations: the "Rewind history" intuition [item 1, p. 93], the "Assume independence" intuition [item 2, p. 93], and, finally, the "No forgetting" intuition underlying Stalnaker's [1998] disagreement with Aumann [1995]. Together with the discussion of the condition of past predominance from Thomason and Gupta [1981] and the condition of identity of overlapping menus [item

1, p. 100], these definitions highlight the subtle issues that arise when merging a logic of counterfactuals with a logic of branching time and agency.

One of the most pressing directions for future research is a sound and complete axiomatization of rewind (resp. independence) models with respect to our full language. We are currently working on a sound and complete axiomatization of ALD_n frames in a language without counterfactuals [cf. Theorem 4.2.7 and Conjecture 4.2.8]. As we mentioned in Section 4.2.2, we expect the proof of completeness over $\mathcal{L}_{\mathsf{ALD}_n}$ to follow the same pattern as the proof of completeness for the logic ALO_n [Appendix A.1] – with some extra steps needed to take care of the Y modality. For our full language $\mathcal{L}^{\Box\!\!\rightarrow}_{\mathsf{ALD}_n}$, we identified some core validities [Propositions 4.3.8 and 4.3.9] and an interesting formula that distinguishes rewind and independence models [Proposition 4.3.11]. Since our definitions of similarity [Definitions 4.3.3, 4.3.4, and 4.4.4] involve counting (deviant) actions, we expect that a complete axiomatization will require an extension of our language. An additional source of complexity not to be overlooked derives from the fact that counterfactuals quantify over *instants*.

Another direction for future research is to explore applications of the logical framework developed in this chapter. Branching-time logics with both agency operators and counterfactuals are a powerful tool to reason about complex social interactions. In particular, as we argued at the end of Chapter 3, logics of this sort seem to be necessary to provide an analysis of notions like causality and responsibility covering cases in which multiple agents act over time. In addition, the discussion in Section 4.4 suggests that a STIT logic with counterfactuals may be fruitfully used to incorporate strategic reasoning in STIT, thus advancing recent research connecting STIT and game-theory [see, e.g., Ciuni and Horty, 2014; Kooi and Tamminga, 2008; Tamminga, 2013; Turrini, 2012].

As we have emphasized above, the aim of this chapter was to investigate the *truth conditions* of choice-driven counterfactuals, not what such counterfactuals express about the knowledge and beliefs of the agents (or even the modeler). In the next Chapter 5, we will shift focus and explore how to model the *mental process* that we use when we evaluate such counterfactuals, studying its structure, logic, and epistemic value.

Chapter 5

Counterfactuals grounded in voluntary imagination

At the end of Chapter 3 we suggested that the formulation and assessment of responsibility judgments centers around two issues related to counterfactual reasoning: (1) identifying the truth conditions of choice-driven counterfactuals, and (2) investigating the structure and epistemic value of the cognitive activity underlying their evaluation. The first issue was the subject of Chapter 4. In the present chapter we aim at addressing the second issue.

There is a wide agreement in cognitive psychology [Byrne, 2005; Byrne and Girotto, 2009], epistemology [Yablo, 1993; Chalmers, 2002; Williamson, 2007], and the philosophy of language [Stalnaker, 1968; Evans and Over, 2004] that we evaluate counterfactuals by relying on a specific form of *imagination*, labeled, for reasons that will become clear soon, *reality-oriented mental simulation* (ROMS). This is the episodic activity of simulating alternatives to reality in our mind and investigate what would happen if they were realized. Here we will focus on *propositional* imagination: one imagines that one jumps a stream, that Obama is tall and thin, that Beth does not plant a device in Alice's brain, that David leaves the game after the first round. We will provide a logical model of propositional imagination as ROMS and consider three connected questions concerning it:

(I) *What is the logic of such an activity*? As persuasively argued by Byrne [2005], exercises of imagination as ROMS must have *some* logic: some things follow from the envisaging of a hypothesis, some do not. For instance, imagination seems to obey logical rules like *Conjunction Commutation* (try to imagine that Obama is tall and thin without imagining that he is thin and tall) and *Elimination* (try to imagine that Obama is tall and thin without imagining that he is tall). The converse, *Conjunction Introduction*, is more controversial: does imagining that A and imagining that B entail imagining that A and B together? Even more controversial are principles like *Closure under Logical Consequence*, at least for non-logically omniscient agents like us: when we imagine that A we do not imagine all of A's logical consequences, and we certainly do not imagine arbitrary

logical truths whenever we engage in an act of ROMS.

(II) *What is the relation between imagination and knowledge*? Connected to question (I) is the issue of how to reconcile imagination's apparent arbitrariness with its having some epistemic value: if imagination is arbitrary escape from reality, how can we get *knowledge through imagination*?[1] Imagination seems to be voluntary in ways contrasting states, like belief, are not: one can easily imagine that all of Amsterdam is painted blue, while one can hardly make oneself believe it, given overwhelming contrary evidence. But if, given some input, one can imagine anything one wills (pending issues of imaginative resistance, see Gendler [2000]), then imagination cannot lead to knowledge.

(III) *What is voluntary in a mental simulation, and what is not*? A promising line of response to question (II) relies on the idea that not *everything* in a ROMS is voluntary, and that the involuntary component often suffices to ensure new, reliably formed, and true beliefs [Williamson, 2016; Langland-Hassan, 2016]. But how to tell the two components apart? The distinction between voluntary and involuntary mental processes is a conundrum in itself, but it seems to make intuitive sense and plays a key role in some mainstream views in cognitive psychology.[2]

The aim of this chapter is to start addressing the above-mentioned questions. We do this in two steps. First, building on literature in cognitive psychology and philosophy, we lay out a general characterization of imagination as ROMS, identifying voluntary and involuntary components. Second, combining techniques from epistemic logic, STIT logic, and subject matter semantics, we design a logic of imagination as ROMS modeling the general characterization of ROMS in order to study the logic and epistemic value of such mental activity.

Outline. We begin, in Section 5.1, by proposing a list of core features of ROMS. In Section 5.2, we introduce our logic of voluntary imagination (VI), which is based on imagination operators that, as we will show, model the identified core features. We present the syntax in Section 5.2.1 and gradually introduce the semantics in the two sections after that: we first define *topic models* [Section 5.2.2] and then build full VI *models* [Section 5.2.3]. We provide a sound and complete axiomatization of the logic VI in Section 5.2.4 and discuss the relation between VI and the logic of imagination $\mathbf{I}^*$ from Berto [2018] and Giordani [2019] in Section 5.2.5. In Section 5.3 we show how our framework allows us to address the three questions concerning imagination as ROMS highlighted above. Section 5.4 concludes.

This chapter is based on Canavotto et al. [2020].

[1] In a recently edited collection bearing this title, Kind and Kung [2016] label this the "puzzle of imaginative use."

[2] For example, dual process theories of thought [Kahneman and Tversky, 1984; Stanovich and West, 2000; Evans and Over, 2004] distinguish between "System 1" and "System 2" processes on the basis of the former being largely automatic, the latter having to be activated and carried out by voluntarily overriding the former's workings, and by paying a cognitive cost.

5.1 Features of imagination as ROMS

The following list of features modifies and complements the one considered in Berto [2018]. In particular, the first and the last features are left unexplored in Berto [2018]. The items in the list have been proposed by researchers on imagination, mental simulation, and pretense, both in philosophy and in cognitive psychology. We will therefore refer to a number of works in both disciplines.

Feature 1. *Imagination is agentive and episodic.*

There is a general agreement that acts of imagination as ROMS are started voluntarily by agents having a number of options given by the situation they are in, their background knowledge and beliefs, and their cognitive abilities [Nichols and Stich, 2003]. We decide to engage in one such act, carry it out for a while, often by controlling some aspects of what we imagine, and stop after some time. This has suggested to some authors [Wansing, 2017] to use STIT logic to model *agents who voluntarily imagine something* (where the stress is on the imagined thing). According to this approach, an agent voluntarily imagines, say, that Alice did not shoot Zac when he sees to it that he imagines Alice not doing such thing. Our concern will be to model *agents who voluntarily engage in certain imagination acts and decide how to carry them out* (where the stress is on the acts themselves): we aim at modeling what it is that an agent does voluntarily when he voluntarily imagines that Alice did not shoot Zac.

Modeling feature 1. For simplicity, we work in a single-agent setting. Taking inspiration from Kripke semantics for STIT [see Sections 2.2.1 and 2.2.3], we start from a set W of possible states and an equivalence relation $R_{\Box}$ over W. Every equivalence class of $R_{\Box}$ represents a situation (a *moment*, in the terminology of STIT) in which the agent can decide to engage in different acts of imagination. But, rather than representing these different acts by partitioning moments into choice-cells, we will use single states instead.[3] The idea is that two states are $R_{\Box}$-equivalent (or *moment-equivalent*) when they are exactly like each other, except possibly for the fact that the agent carries out different imagination exercises at them. We can then say that the agent has the option to conduct a certain ROMS in a certain way at a state w when there is a state w' that is moment-equivalent to w at which she carries out that ROMS in that way. So, equivalent states keep track of the options the agent has in conducting a ROMS. In the syntax, we will have a modal operator $\Box$ to talk about what is settled true, i.e., what the agent cannot control in a ROMS (its dual $\Diamond$ will allow us to talk about what the agent can control in it). For uniformity with the previous chapters, we will refer to this operator as the operator of *historical necessity*.

[3] A similar idea can be found in Giordani [2018].

Feature 2. *Imagination acts have a deliberate starting point, given by an input.*
Such input is up to us. In their model of mental simulation, Nichols and Stich [2003, p. 24] have "an initial premise or set of premises, which are the basic assumptions about what is to be pretended." This may be made up by the agent when engaging in predictions, e.g., when we guess what would happen if something were the case; or it may be taken on board via an external instruction, e.g., when we read a novel and take the explicit text as our input, or when we evaluate a conditional and start by taking the antecedent as input. Suppositional theories of conditionals in psychology as well as philosophy connect this to the so-called Ramsey test [see Evans and Over, 2004, pp. 21-25].

Modeling feature 2. In our formalism below, we will represent the explicit input as directly expressed, in the syntax, by formulas indexing imagination operators and featuring, in the semantics, as arguments of functions that model the imagined content. We will say more about these functions when we discuss Feature 3 and Feature 5.

Feature 3. *We integrate the explicit input with background information that we import, contextually, depending on what we know or believe.*
The importance of background knowledge and beliefs in suppositional thinking is increasingly acknowledged in the psychology of reasoning [Oaksford and Chater, 2010]. Once the input of an act of mental simulation is in, Nichols and Stich [2003, pp. 26-28] claim, "children and adults elaborate the pretend scenarios in ways that are not inferential at all," filling in the explicit instruction with "an increasingly detailed description of what the world would be like if the initiating representation were true." When we imagine Watson talking with Holmes while walking through the streets of London, we represent Watson dressed as a nineteenth Century gentleman, not as an astronaut. The text of the relevant novel need not say anything explicitly on how Watson is dressed, nor do we infer this from the explicit content via sheer logic. Rather, we import such information into the represented situation, based on what we know: we know that the story takes place in Victorian London and we assume, lacking information to the contrary from the text, that Watson is dressed as we know gentlemen were dressed at the time. We perform some *minimal alteration* to how we know or believe the world to be, or to have been, compatible with the initial explicit input, in a process somewhat similar to belief revision [Alchourrón et al., 1985], whereby we perform a minimal change of our belief system needed to accommodate new information.

A key issue here is: Is this integration process voluntary or not? Some authors seem to agree that the involuntary component of imagination as ROMS comes into play exactly here. As Williamson [2016, p. 116] has it:

> Think of a hunter who finds his way obstructed by a mountain stream rushing between the rocks. [...] How should he try to determine

> whether he would succeed [if he jumped]? [...] One imagines oneself trying. If one then imagines oneself succeeding, one judges that if one tried, one would succeed. If instead one imagines oneself failing, one judges that if one tried, one would fail. [...] When the hunter makes himself imagine trying to jump the stream, his imagination operates in voluntary mode. But he neither makes himself imagine succeeding nor makes himself imagine failing. Rather, having forced the initial conditions, he lets the rest of the imaginative exercise unfold without further interference. For that remainder, his imagination operates in involuntary mode. He imagines the antecedent of the conditional voluntarily, the consequent involuntarily.

Similarly, Langland-Hassan [2016] distinguishes between "guiding chosen" imaginings, that is, "top down intentions [that] are key to initiating an imagining," and "lateral constraints [that] govern how it then unfolds," which seem to operate in involuntary mode. If the additional details are borrowed from our knowledge or belief base, as Van Leeuwen [2016] and Nichols and Stich [2003] have it,[4] this makes sense: for if beliefs are often formed and managed in largely involuntary mode, it seems plausible for their importation to be essentially involuntary. Some research in cognitive psychology seems to support the view that imagination allows automatic, involuntary access to the knowledge deposited in implicit (long-term) memory, and that the results of imaginative exercises can themselves alter such memory [Kosslyn and Moulton, 2009].

Modeling feature 3. It is natural to represent this integration of the initial input via background beliefs and knowledge by using modal operators that work as *variably strict* quantifiers over states: the input will play a role similar to a variably strict conditional antecedent, as per the mainstream possible worlds semantics for counterfactuals due to Stalnaker [1968] and Lewis [1973a]. In the semantics we will have a function f_{in} that selects, for every input φ and state w, the states that are consistent both with input φ and the relevant background beliefs of the agent at w (for short, we will call them *the closest φ-states to w*). This squares with the aforementioned psychological insight that we evaluate conditionals by mentally representing the antecedent, developing it in our imagination, and seeing whether the consequent would, in some sense, follow [Oaksford and Chater, 2010].

Feature 4. *Imagination has topicality and relevance constraints.*

We do not indiscriminately import unrelated contents into the imagined scenarios. As Kind [2016, p. 153] has it, "[We require] that the world be imagined as it is *in all relevant respects*." This is key to distinguishing imagination as ROMS from free-floating mental wandering. We know that Amsterdam hosts plenty of bikes,

[4] Nichols and Stich [2003] have a cognitive "belief box," from which contents are taken and imported into the mental simulation.

but this is immaterial to our imagining Watson and Holmes' adventures from Doyle's novels, insofar as Amsterdam and its bikes are not involved in them. So we will not import knowledge of this kind, even when it is perfectly consistent with the explicit input.

Modeling feature 4. This *topic-sensitivity* of ROMS will be modeled by imposing topic-preservation constraints on the outputs of imaginative exercises, which secure their complete relevance with respect to the starting point given by the explicit input. Our logical models will feature (formal representations of) *topics* to do the job (we will see what these are in due time).

Feature 5. *The content of ROMS is goal-driven and question-based.*

Together with Feature 1, this item was left unexplored in Berto [2018], but it is of the greatest importance. Acts of ROMS have a *goal* [Fraude-Koivisto et al., 2009], which can be understood via the question, or issue, the agent performing them aims to answer. This is crucial because the same agent, with the same background beliefs, and given the same input, can focus on different things depending on the goal of the exercise, and thereby end up imagining quite different scenarios. Let us illustrate this point by elaborating on Williamson's [2016] example.

5.1.1. Example. We are planning a hike in the Alps and we are informed that there is a stream that crosses the route. We start imagining what would happen if we jump. The input of our imagining is *I jump the stream (in the Alps)* and the issue is *Will I make it to the other side?*. With this issue in mind, we focus on, e.g., our weight and the conditions of the ground, while we leave behind aspects less relevant for the question, such as the kind of noises we would make while jumping. But, if the issue were *Will I scare the cattle on the other side?*, we would rather focus on the kind of noises we would make while jumping.

On the psychological side, it seems that the choice of the goal to be pursued in an act of ROMS – hence of which aspects of the imagined scenario are relevant – is voluntarily: the question *Will I make it to the other side?* is what we set out to answer in the act. At the same time, the specific way in which the relevant aspects are represented seems to be partly involuntary and partly voluntary: while some aspects are fixed by our background knowledge and beliefs (as when we believe that we weigh 55kg), others will be up to us (as when we opt for being optimistic about the weather and imagine a dry, sunny day).

Modeling feature 5. The goal of a ROMS exercise determines the salience of certain traits related to the input, and the fading in the background of others. This phenomenon can be represented by having any input φ and goal τ determine a partition of W (we will call it the *τ-partition for φ*). With Example 5.1.1, given input *I jump the stream (in the Alps)* and issue *Will I make it to the other side?*, states that are like each other with respect to facts relevant to the input (as the

kind of landscape surrounding the stream) and to the issue (as our weight and the conditions of the ground) will belong to the same cell. For instance, the following states w_1 and w_2 will belong to the same cell:

- at w_1 the stream is surrounded by nature, we weigh 55kg, the ground by the stream is dry, we jump silently, a stream in the Shenandoah National Park is surrounded by nature;
- at w_2 the stream is surrounded by nature, we weigh 55kg, the ground by the stream is dry, we jump noisily, the stream in the Shenandoah National Park is surrounded by skyscrapers.

Yet, if the input were *I jump the stream (in the Shenandoah National Park)* or the issue were *Will I scare the cattle?*, w_1 and w_2 would belong to different cells.

We will represent the goals of acts of imagination by means of *topics*, which have been naturally linked to partitions of the set of possible states [see Section 5.2.2]. In addition, we will model the (partly voluntary) process of specification of the relevant traits determined by the input and topic of a ROMS by means of a selection function f_{top}. This picks, for any state w, input φ, and topic τ, the closest φ-states to w where the traits determined by φ and τ are specified as the agent specifies them at w. So, in Example 5.1.1, at a state where we opt for being optimistic about the weather, f_{top} selects the closest states where we jump the stream and the ground is dry. Another way to think of f_{top} is that it selects the *cells* in the partition determined by the input and topic where the ground is dry. In the syntax, we will have both modal operators to talk about what is necessarily the case in the cell of this partition that includes the actual state and modal operators to talk about what is necessarily the case in the cells of the partition that are selected by the agent at the actual state.

General structure of ROMS. Let us wrap up with a general description of a ROMS episode. When the input comes in, two things happen: (1) in involuntary mode, we integrate the input with relevant background beliefs; (2) in voluntary mode, we set the goal of the exercise and determine which traits in the imagined scenario are salient. We will call the (overall, partly voluntary) specification of these traits *specified imagined scenario*, to distinguish it form the *basic imagined scenario* that results from integration (1), independently from goal-setting (2).[5] Next, the specified imagine scenario unfolds in involuntary mode, generating new beliefs that we may use in subsequent imagination acts, like the belief that, if we try to jump the stream and the ground is dry, we will make it to the other side.

Figure 5.1 is a functional representation of the general structure of imagination as ROMS. Darker gray, lighter gray, and white boxes represent involuntary,

[5]It is easy to see that (1) is independent of (2). In Example 5.1.1, when we receive the input *I jump the stream (in the Alps)*, we automatically picture the stream surrounded by nature, not by skyscrapers, regardless of the question we aim to answer.

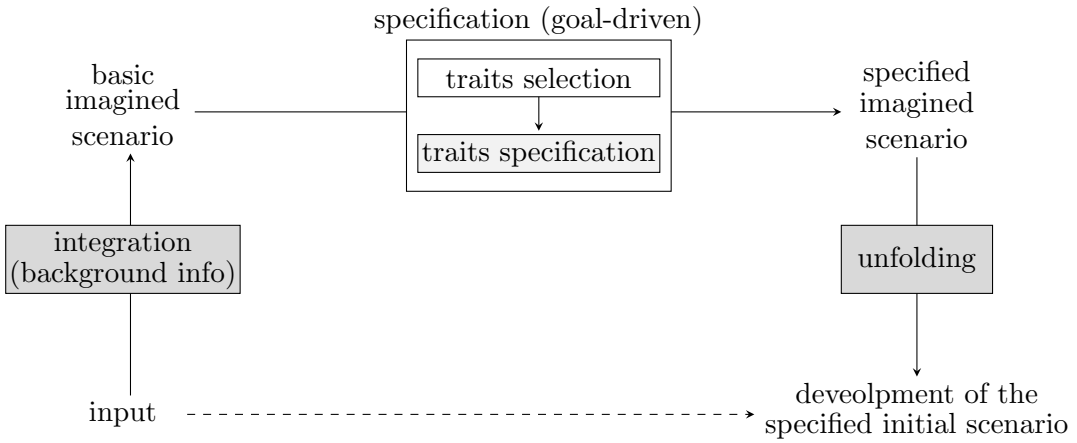


Figure 5.1: General structure of ROMS

partly voluntary, and voluntary processes respectively. The goal-driven process of specification of the basic initial scenario is decomposed into the processes of selection and specification of relevant traits. The arrow from the input to the development of the specified initial scenario is dashed to highlight that the overall process from the former to the latter is a mediated, rather than a direct, one. With respect to this general schema, the logic we present in the next Section 5.2 aims at modeling, first, the processes of integration of the input (as per Feature 2 and Feature 4) and specification of the basic imagined scenario (as per Feature 4 and Feature 5) and, second, the control the agent can exercise on the two processes (as per Feature 1). The logic gives only a basic representation of involuntary unfolding in terms of what is necessary, or logically follows, given the imagined initial conditions and the topic of the input. In this regard, it makes sense to say that the specified imagined scenario is *unfolded in time*, as episodes of ROMS can last for an amount of time and can involve representing actions and events, which themselves evolve in time. We have investigated different ways to model this kind of unfolding in STIT semantics in Chapter 4. We conjecture that, in the present framework, this may be best modeled by dynamic operators whose semantics is given in terms of model-transformations [Baltag et al., 1998; van Benthem, 2011; van Ditmarsch et al., 2008], leaving this to future work.

5.2 The logic of voluntary imagination VI

The logic of voluntary imagination VI extends the logic of imagination $\mathbf{I}^*$ introduced by Berto [2018] and axiomatized by Giordani [2019]. In this section, we present its syntax and semantics, provide a sound and complete axiomatization,

and prove that VI is a conservative extension of $\mathbf{I}^*$.

5.2.1 Syntax

The language $\mathcal{L}_{\mathsf{VI}}$ of the logic VI is built from a countable set of propositional variables and a countable set of names of topics. It includes five types of modalities: the *universal modality* A; the modality $\Box$ of *historical necessity*; a family of *imagination modalities* indexed by an input (modalities like $[im_\varphi]$); a family of *topic-driven imagination modalities* indexed by an input and a topic (modalities like $[im_{(\varphi,\tau)}]$); and a family of modalities of *input-and-topic necessity* indexed by an input and a topic (modalities like $[nec_{(\varphi,\tau)}]$).

5.2.1. Definition (Syntax of $\mathcal{L}_{\mathsf{VI}}$). Given a countable set T of names for topics and a countable set *Prop* of propositional variables, the set of formulas of $\mathcal{L}_{\mathsf{VI}}$, also denoted with $\mathcal{L}_{\mathsf{VI}}$, is generated by the following grammar:

$$\varphi := p \mid \neg\varphi \mid (\varphi \wedge \varphi) \mid \mathsf{A}\varphi \mid \Box\varphi \mid [im_\varphi]\varphi \mid [im_{(\varphi,\tau)}]\varphi \mid [nec_{(\varphi,\tau)}]\varphi$$

where $p \in \mathit{Prop}$ and $\tau \in T$. The abbreviations for the other Boolean connectives are standard. In addition, we use the following abbreviations for the dual modalities: $\mathsf{E}\varphi := \neg\mathsf{A}\neg\varphi$, $\Diamond\varphi := \neg\Box\neg\varphi$, $\langle im_\varphi\rangle\,\psi := \neg[im_\varphi]\neg\psi$, $\langle im_{(\varphi,\tau)}\rangle\,\psi := \neg[im_{(\varphi,\tau)}]\neg\psi$, $\langle nec_{(\varphi,\tau)}\rangle\,\psi := \neg[nec_{(\varphi,\tau)}]\neg\psi$. We will adopt the usual conventions for the elimination of parentheses.

The intended interpretation of the modal formulas is as follows. As usual, $\mathsf{A}\varphi$ means that φ is true at all possible states and $\Box\varphi$ means that φ is settled true or historically necessary at the present moment. When φ is a formula about the agent's imaginings, $\Box\varphi$ means that the agent has no control on φ. A formula like $[im_\varphi]\psi$ is read "given input φ, the agent imagines that ψ (independently of the selected topic)." We take $[im_\varphi]\psi$ to be true at a state w when ψ is true at all the closest φ-states to w *and* on-topic relative to φ (we will see what this means in a moment). Imagination modalities like $[im_\varphi]$ correspond to the original imagination operators proposed by Berto [2018]. With respect to $[im_\varphi]\psi$, formulas like $[im_{(\varphi,\tau)}]\psi$ also take topics into account: $[im_{(\varphi,\tau)}]\psi$ is read "given input φ *and the selected specification of topic* τ, the agent imagines that ψ."[6] This formula is taken to be true at a state w when ψ is on-topic relative to φ *and* true throughout the cells of the τ-partition for φ selected by the agent at w. Finally, $[nec_{(\varphi,\tau)}]\psi$ means that ψ is true at all states in the cell of the τ-partition for φ to which the present state belongs.

When the semantics for $\mathcal{L}_{\mathsf{VI}}$ is given, we will see that the validities involving $[im_\varphi]$ and $[im_{(\varphi,\tau)}]$ will speak to our first issue – the question of the logic of

[6] We write "the selected specification of topic τ" to abbreviate "the selected specification of the salient traits determined by the input and topic τ."

imagination as ROMS. As for the other two issues – the question of the relation between ROMS and knowledge and the question of the voluntary and involuntary components of ROMS – we need a preliminary definition.

5.2.2. Definition. Where $\varphi \in \mathcal{L}_{\mathsf{VI}}$ and $Var(\varphi) \subseteq Prop$ is the set of propositional variables occurring in φ,[7] we set:

$$\overline{\varphi} := \bigwedge_{p \in Var(\varphi)} (p \vee \neg p)$$

So, $\overline{\varphi}$ is the conjunction of all the instances of the principle of excluded middle given by propositional variables occurring in φ. It is evident that $Var(\overline{\varphi}) = Var(\varphi)$ and that, for any propositional variable p, $\overline{p} = p \vee \neg p$. As it will become clear after Definition 5.2.5 below, using this kind of formulas – rather than the more familiar constant $\top$ – is essential, in our setting, to express topic inclusion. This, in turn, will allow us to use $\langle im_{\varphi} \rangle \overline{\varphi}$ to express that the content imagined by the agent given input φ is not empty and, similarly, $\langle im_{(\varphi,\tau)} \rangle \overline{\varphi}$ to express that the content imagined by the agent given input φ and the selected specification of topic τ is not empty. This justifies reading $\langle im_{\varphi} \rangle \overline{\varphi}$ as "the agent is engaged in an act of imagination based on input φ" and $\langle im_{(\varphi,\tau)} \rangle \overline{\varphi}$ as "the agent is processing input φ in light of topic τ." The combination of the latter two formulas with the modal $\Box$ will be key to address our second and third issues.

5.2.2 Topics and topic models

As in Berto [2018], the semantics we propose includes, besides the usual set W of possible states, a set of topics, T.[8] Everyone is familiar with the former, while we need to say something by way of introduction to the latter, which have already been invoked above.

One can understand topics as somewhat similar to Lewisian or Yablovian *subject matters* from aboutness theory [Lewis, 1988; Yablo, 2014]. Aboutness is "the relation that meaningful items bear to whatever it is that they are *on* or *of* or that they *address* or *concern*" [Yablo, 2014, p. 1] – their subject matters, or topics. In works such as Lewis' or Yablo's, these are understood in relation to questions: the subject matter or topic of sentence S in context c can be linked to the question(s) S can be an answer to in c. When the topic at issue is *the number of stars*, the corresponding question can be *How many stars are there?*.

[7] $Var(\varphi)$ is defined recursively in the usual way:

- for $\varphi := p$: $Var(\varphi) = \{p\}$;
- for $\varphi := \neg\psi$: $Var(\varphi) = Var(\psi)$;
- for $\varphi := \psi \wedge \chi$, $\varphi := [im_{\psi}]\chi$, $\varphi := [im_{(\psi,\tau)}]\chi$, $\varphi := [nec_{(\psi,\tau)}]\chi$: $Var(\varphi) = Var(\psi) \cup Var(\chi)$.

[8] For the sake of simplicity, we will abuse notation and treat the set T both as a set of names of topics in the syntax and the named topics in the semantics.

This determines a partition on W. Two worlds end up in the same cell when they give the same answer to the question. So, all zero-star worlds end up in one cell, all one-star worlds end up in another, and so on.[9]

Being about stuff – having a topic – is not only a feature of linguistic representations, but also, and perhaps more fundamentally, of mental ones. In particular, imagining is about stuff. In our setting, topics are what the mental states of imaginative agents are about. The topic of an imaginative exercise is given via its specific purpose, which can also be understood via a question. Recall Example 5.1.1 above. When the goal is predicting whether we will make it to the other side if we jump the stream, the question will be, *Will I make it to the other side?*. The possible answers will depend on the specification of certain salient traits, which can also be spelled out as questions, like *How much do I weigh?* or *What are the conditions of the ground?*. The set of all these questions determines a partition on W: two worlds end up in the same cell when they give the same answer to all the questions in the set – when they agree on our weight, the conditions of the ground, and so on. So, imaginative exercises have a topic (*Will I make it to the other side?*), which determines a set of questions (*How much do I weigh?*, *What are the conditions of the ground?*, etc.), which, together with facts relevant to the input (*Is the stream surrounded by nature?*), determine a partition of the set of possible states. We will see that the selected answers, corresponding to a union of cells in the corresponding partition, are captured by the modality $[im_{(\varphi,\tau)}]\psi$.

But *how* are topics like? What is their nature and structure? We do not need to say too much on this for our logical purposes, except that we need a recursion on them allowing us to come up with a compositional semantics for our language. Luckily, there is a natural mereology of topics at the sentential level: what a sentence is about can be (properly) included in what another one is about [Yablo, 2014, Section 2.3; Fine, 2016, Sections 3-5]. Topics may be merged into wholes that inherit the proper features from the parts [Yablo, 2014, Section 3.2]. The Boolean connectives are topic-transparent – they have no subject matter of their own: φ has the same topic as $\neg\varphi$ ("Snow is not white" is about the color of snow, or how snow is like, or snow's whiteness, etc., just as "Snow is white"); $\varphi \wedge \psi$ and $\varphi \vee \psi$ have the same topic, i.e., a fusion of the subject matter of φ and that of ψ ("Obama is tall and thin" is about Obama's heights and figure, just as "Obama is tall or thin"). We extend transparency to our modals in a straightforward[10] way, and come up with the following.

[9] Yablo [2014] proposes to generalize and replace partitions with divisions, which allow worlds to be in more than one cell, for a question, e.g., *Where is a nice place to eat in Amsterdam?*, can have more than one good answer. We will not complicate things accordingly in our framework.

[10] But not uncontroversial. "Obama is tall and thin" and "Tom imagines that Obama is tall and thin" seem to have different topics: only the latter is about Tom's mental states. Fortunately, for our logical purposes the difference only matters when one considers nested or higher-order imaginings (one imagines that one imagines that), which are beyond the scope of this dissertation.

5.2.3. Definition (Topic model for $\mathcal{L}_{\mathsf{VI}}$). A topic model for $\mathcal{L}_{\mathsf{VI}}$ is a tuple $\mathcal{T} = \langle T, \oplus, t\rangle$, where $T \neq \varnothing$ is a set of topics, $\oplus : T \times T \to T$ is a *fusion operation*, and $t : Prop \to T$ is a *topic function* assigning a topic to every propositional variable in $Prop$. The fusion operation $\oplus$ is required to satisfy the following properties, for all $\tau_1, \tau_2, \tau_3 \in T$:

- *Idempotency*: $\tau_1 \oplus \tau_1 = \tau_1$.
- *Commutativity*: $\tau_1 \oplus \tau_2 = \tau_2 \oplus \tau_1$.
- *Associativity*: $\tau_1 \oplus (\tau_2 \oplus \tau_3) = (\tau_1 \oplus \tau_2) \oplus \tau_3$.

The topic function t is extended to the whole $\mathcal{L}_{\mathsf{VI}}$ as follows. For any $\varphi \in \mathcal{L}_{\mathsf{VI}}$, if $Var(\varphi) = \{p_1, \ldots, p_n\}$, then the topic of φ is:

$$t(\varphi) = t(p_1) \oplus \ldots \oplus t(p_n)$$

The idea behind the extension of the topic function t to the whole $\mathcal{L}_{\mathsf{VI}}$ is that a formula is about whatever its atomic components taken together are about. It follows immediately from this that, for all $\varphi \in \mathcal{L}_{\mathsf{VI}}$, $t(\varphi) = t(\overline{\varphi})$ and that the Boolean connectives as well as the modal operators of $\mathcal{L}_{\mathsf{VI}}$ are topic-transparent. That is, the following hold for all $\varphi, \psi \in \mathcal{L}_{\mathsf{VI}}$ and $\tau \in T$:

1. $t(\neg\varphi) = t(\mathsf{A}\varphi) = t(\Box\varphi) = t(\varphi)$.
2. $t(\varphi \wedge \psi) = t([im_{\varphi}]\psi) = t([im_{(\varphi,\tau)}]\psi) = t([nec_{(\varphi,\tau)}]\psi) = t(\varphi) \oplus t(\psi)$.

We define a relation $\sqsubseteq \subseteq T \times T$ of *topic inclusion* in the standard way by setting, for all $\tau_1, \tau_2 \in T$:

$$\tau_1 \sqsubseteq \tau_2 \text{ if and only if } \tau_1 \oplus \tau_2 = \tau_2$$

Thus, $\langle T, \oplus\rangle$ is a *join semilattice* and $\langle T, \sqsubseteq\rangle$ a *partially ordered set*. Intuitively, given $\varphi, \psi \in \mathcal{L}_{\mathsf{VI}}$, $t(\psi) \sqsubseteq t(\varphi)$ holds when ψ *is on-topic relative to* φ. Let us explain. In terms of questions, $t(\varphi)$ can be viewed as the most specific question φ can be an answer to.[11] Under this interpretation, $t(\psi) \sqsubseteq t(\varphi)$ is the case when $t(\psi)$ corresponds to a question whose answers are disjunctions of answers to the question associated with $t(\varphi)$. When this happens, the topic of ψ allows no alien element with respect to φ, that is, ψ is about stuff φ is also about – in this sense, ψ is "on-topic" relative to φ. It can be easily checked that, if $Var(\psi) \subseteq Var(\varphi)$ then $t(\psi) \sqsubseteq t(\varphi)$. So, "Watson talks to Holmes" is on-topic relative to "Watson talks to Holmes and Irene Adler leaves England," while "Watson talks to Holmes and our friend was late because of a mascara accident" may not be.

[11] An extensional implementation of this idea can be found in Lewis [1988, pp. 161-163]. There, the topic of a proposition on the 17th Century corresponds to the most specific question concerning that Century. The question is modeled as a partition whose cells stand for the most specific exclusive descriptions of the 17th Century – i.e., the possible answers to the question. Any other (less specific) proposition on the 17th Century is thus a union of cells in that partition.

5.2.3 Semantics

A VI model M consists of a number of components. First, a topic model $\mathcal{T}$ that fixes the topic of every formula of $\mathcal{L}_{\mathsf{VI}}$ and a set W of possible states.[12] Next, an relation $R_{\Box}$ of moment-equivalence and a function $\sim$ assigning to every possible input φ and topic τ an equivalence relation $\sim_{(\varphi,\tau)}$ that determines the τ-partition for φ. Finally, two selection functions f_{in} and f_{top}. The former takes an input φ and a state w and selects the closest φ-states to w; the latter takes an input φ, a topic τ, and a state w and selects the states in the cells of the τ-partition for φ that the agent opts for at w. As we will also emphasize later on, both selections can be empty.

5.2.4. Definition (VI model). A VI model is a tuple $\langle \mathcal{T}, W, R_{\Box}, \sim, f_{in}, f_{top}, \nu \rangle$, where $\mathcal{T} = \langle T, \oplus, t \rangle$ is a topic model, $W \neq \varnothing$ is a set of possible states, $R_{\Box} \subseteq W \times W$ is an equivalence relation, and $\nu : \mathit{Prop} \to 2^W$ is a valuation function. In addition,

- $\sim$: $\mathcal{L}_{\mathsf{VI}} \times T \to 2^{W \times W}$ assigns an equivalence relation $\sim_{(\varphi,\tau)} \subseteq W \times W$ to every $(\varphi, \tau) \in \mathcal{L}_{\mathsf{VI}} \times T$;
- $f_{in} : \mathcal{L}_{\mathsf{VI}} \times W \to 2^W$ assigns to every $(\varphi, w) \in \mathcal{L}_{\mathsf{VI}} \times W$ the set $f_{in}(\varphi, w)$ of closest φ-states to w;
- $f_{top} : \mathcal{L}_{\mathsf{VI}} \times T \times W \to 2^W$ assigns to every $(\varphi, \tau, w) \in \mathcal{L}_{\mathsf{VI}} \times T \times W$ the set $f_{top}(\varphi, \tau, w)$ of states in the cells of the τ-partition for φ selected at w.

The elements of VI models are required to satisfy the following conditions: for all $\varphi \in \mathcal{L}_{\mathsf{VI}}$, $\tau, \tau_1, \tau_2 \in T$, and $w, v, v' \in W$:

1. *No choice of the basic imagined scenario*: if $wR_{\Box}v$, then $f_{in}(\varphi, w) = f_{in}(\varphi, v)$.
2. *Specification of the basic imagined scenario*: $f_{top}(\varphi, \tau, w) \subseteq f_{in}(\varphi, w)$.
3. *No parting of indistinguishable states*: if $v \in f_{top}(\varphi, \tau, w)$ and $v \sim_{(\varphi,\tau)} v'$, then $v' \in f_{top}(\varphi, \tau, w)$.
4. *Ability to select answers*: if $v \in f_{in}(\varphi, w)$, then there is $w' \in W$ such that $wR_{\Box}w'$ and $v \in f_{top}(\varphi, \tau, w')$.

In case you are feeling a bit lost concerning the behavior of the elements of a VI model, let us go back to Example 5.1.1, where we receive input *I jump the stream* and select topic *Will I make it to the other side?*. The example can be

[12] The topic model $\mathcal{T}$ will not be relative to states: every formula will have the same topic at all states of any given VI model. This is a simplification and is not essential to what follows. In Canavotto et al. [2020] we favored generality over simplicity and defined VI models by assigning a topic model to every possible state.

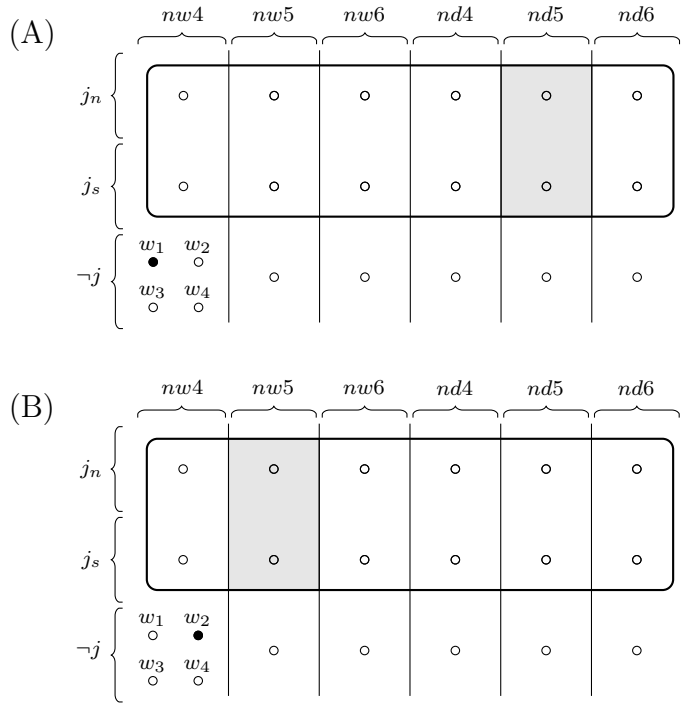


Figure 5.2: Modeling Example 5.1.1, an illustration

modeled as illustrated in Figure 5.2. The labels at the top of Diagrams (A) and (B) are read as follows: n stands for "the stream is surrounded by nature," w for "the ground is wet," d for "the ground is dry," and the numbers $4, 5$, and 6 for "I weigh 54kg," "I weigh 55kg," and "I weigh 56kg" respectively. From top to bottom, the labels on the left of the diagrams mean: "I jump noisily," "I jump silently," and "I do not jump." Circles represent possible states. The actual state (in black) is w_1 in Diagram (A) and w_2 is Diagram (B). These are both states where we do not jump, the stream is surrounded by nature, the ground around it happens to be wet, and we weigh 54kg.

Consider now Diagram (A). When input *I jump the stream* (let us write it as p) comes in, we integrate it with relevant background beliefs, like the belief that the stream is surrounded by nature and the belief that we are fit (i.e., we weigh something between 54kg and 56kg). A set $f_{in}(p, w_1)$ of states (represented by the thick rectangle) is then selected: these are the states consistent with p and with the imported beliefs. Also, the topic *Will I make it to the other side?* (let us write it as τ) is chosen, and so questions like *How much do I weigh?* and *What are the conditions of the ground* become salient. Together with the input, these questions determine a partition of the set of all states, which is modeled by the relation $\sim_{(p,\tau)}$ (this is represented by the vertical lines in the diagram): states

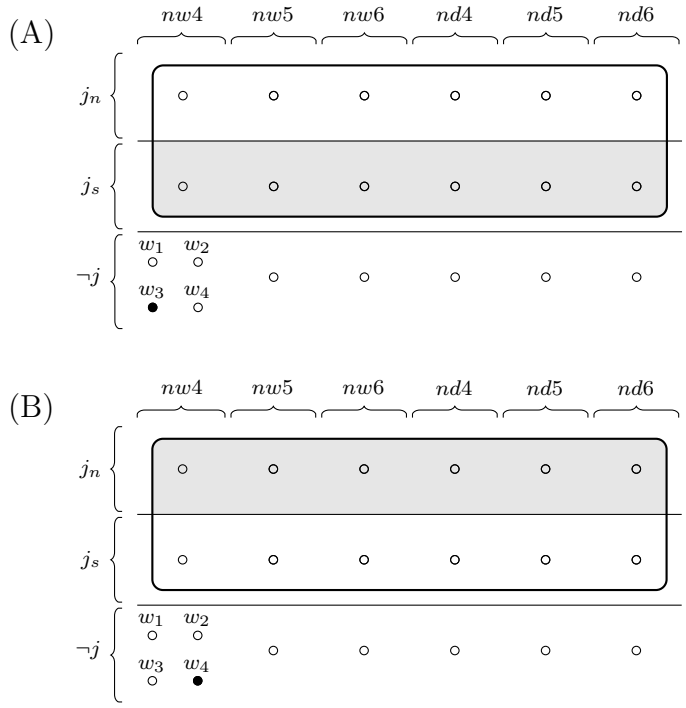


Figure 5.3: Modeling Example 5.1.1, an illustration

that agree on the kind of landscape surrounding the stream, our weight, and the conditions of the ground end up in the same cell. Finally, cells $f_{top}(p, \tau, w_1)$ in this partition are picked out (the gray shaded area): these represent the answers we give to the questions determined by the input and topic, i.e., that we weigh 55kg and the ground is dry.

The upshot is that $f_{in}(\varphi, w_1)$ and $f_{top}(\varphi, \tau, w_1)$ just stand for the *basic imagined scenario* and the *specified imagined scenario* in an act of ROMS (recall the functional representation in Figure 5.1). Function $\sim$ is needed to represent the partition of W induced by the input and topic: without it, we would not be able to connect topics with partitions, and so to explicitly model the selection of cells in a partition.

As for $R_{\Box}$, consider again world w_1 and look at Figure 5.2 (B). Since we have an option to be pessimistic about the weather, there is a state, *viz.* w_2, that is like w_1 in all respects except that, after receiving input p and selecting topic τ, we pick cells where the ground is wet rather than dry. In addition, as illustrated by the two diagrams in Figure 5.3, since we are free to select the topic itself when we receive the input, there are states, *viz.* w_3 and w_4, that are like w_1 in all respects except that, after receiving and integrating the input with relevant background beliefs, we select a different topic, such as *Will I scare the cattle?* (write it as

τ'). This triggers the question *Would I jump noisily?*, which, together with the input, determines the partition represented by the horizontal lines. States w_3 and w_4 keep track of the fact that we can opt to imagine ourselves either as jumping silently ($f_{top}(p, \tau', w_3)$, represented by the gray shaded area in Figure 5.3 (A)) or as jumping noisily ($f_{top}(p, \tau', w_4)$, represented by the gray cell in Figure 5.3 (B)).

Going back to Definition 5.2.4, conditions 1 to 4 ensure that the elements of VI models indeed work as depicted in the figures. Condition 1 (no choice of the basic initial scenario) is based on the idea that, if states w and w' are exactly alike except (perhaps) for what the agent imagines, then w and w' must be alike with respect to the agent's background beliefs. Hence, the set of closest φ-states to w and the set of closest φ-states to w' must be the same. Condition 2 (specification of the basic imagined scenario) ensures that what is imagined given an input and a selected topic is a specification of what is imagined given the input. Condition 3 (no parting of indistinguishable states) roughly says that no answer the agent can give to the questions determined by an input and topic can take apart states that give the same answers to those questions. Technically, this ensures that $f_{top}(\varphi, \tau, w)$ is a union of cells in the partition induced by $\sim_{(\varphi,\tau)}$. Finally, condition 4 (ability to select answers) has it that any state consistent with input φ and the beliefs imported relative to it represents an answer the agent can give to the questions determined by φ together with any topic.

5.2.5. Definition (Semantics for $\mathcal{L}_{\mathsf{VI}}$). Given a VI model M, truth of a formula $\varphi \in \mathcal{L}_{\mathsf{VI}}$ at a world w in M, denoted $M, w \models \varphi$, is defined recursively. Truth of atomic propositions and the Boolean connectives is defined as usual. The remaining cases are as follows:

$M, w \models \mathsf{A}\varphi$	iff	for all $w' \in W$, $M, w' \models \varphi$
$M, w \models \Box\varphi$	iff	for all $w' \in W$, if $wR_{\Box}w'$, then $M, w' \models \varphi$
$M, w \models [im_{\varphi}]\psi$	iff	for all $w' \in W$, if $w' \in f_{in}(\varphi, w)$, then $M, w' \models \psi$, and $t(\psi) \sqsubseteq t(\varphi)$
$M, w \models [im_{(\varphi,\tau)}]\psi$	iff	for all $w' \in W$, if $w' \in f_{top}(\varphi, \tau, w)$, then $M, w' \models \psi$, and $t(\psi) \sqsubseteq t(\varphi)$
$M, w \models [nec_{(\varphi,\tau)}]\psi$	iff	for all $w' \in W$, if $w \sim_{(\varphi,\tau)} w'$, then $M, w' \models \psi$

For any formula $\varphi \in \mathcal{L}_{\mathsf{VI}}$, $[\![\varphi]\!]^M = \{w \in W \mid M, w \models \varphi\}$ is the *truth-set of* φ *in* M. We will omit reference to the model M when it is clear from the context.

Definition 5.2.5 takes into account a number of features of ROMS. The indexing of $[im_{\varphi}]$ and $[im_{(\varphi,\tau)}]$ by φ deals with Feature 2: the starting point of imagination acts comes with the input, deliberately chosen. For $[im_{\varphi}]\psi$ to be true at w we require that two conditions are met:

1. ψ must be true at the closest φ-states to w, where the variability in the sets selected given different inputs captures the variability of beliefs imported,

relative to such inputs, in accordance with Feature 3;[13]

2. ψ must be fully on-topic with respect to the input, thereby capturing the relevance or topicality constraints of a proper act of imagination, in accordance with Feature 4.

For $[im_{(\varphi,\tau)}]\psi$ to be true at w, besides constraint 2, we require ψ to be true throughout the states selected given input φ *and* topic τ at w. The topic is the issue or question addressed in the imaginative exercise and contributes to fix what it is about, in accordance with Feature 5. Definition 5.2.5 also makes evident the difference between $[nec_{(\varphi,\tau)}]$ and $[im_{(\varphi,\tau)}]$: equivalence with respect to inputs and topics is an objective feature of a given state, so the truth of $[nec_{(\varphi,\tau)}]\psi$ is independent of relevance conditions concerning topics. Imaginability with respect to inputs and topics depends on the imaginative acts and choices of the agent, so the truth of $[im_{(\varphi,\tau)}]\psi$ also depends on relevance conditions on topics.

The difference between the three modalities $[im_\varphi]$, $[im_{(\varphi,\tau)}]$, and $[nec_{(\varphi,\tau)}]$ can be illustrated again with Figure 5.2 (A). Let p, τ, d, and w be as before. Also, let r be the sentence that there are plenty of bikes in Amsterdam. Assume that this sentence is true at all states and that, unlike d and w, it is not on topic relative to p (that is, $t(d) \sqsubseteq t(p)$, $t(w) \sqsubseteq t(p)$, but $t(r) \not\sqsubseteq t(p)$). Then, at state w_1, $[im_p](d \vee w)$ is true but $[im_p]r$ is not: although the closest p-states to w_1 are all states where the ground is either dry or wet and there are plenty of bikes in Amsterdam, the latter is not on-topic with respect to p. So, given input p, the agent imagines that the ground is either dry or wet, but not that there are plenty of bikes in Amsterdam. Also, $[im_{(p,\tau)}]d$ is true at w_1 but $[im_p]d$ is not: although d is on-topic relative to p, the beliefs imported when the input is received are not sufficiently strong, *by themselves*, to rule our the possibility that the ground is wet. So, given input p *and* the choice to be optimistic about the weather, the agent imagines that the ground is dry, even if, before making this choice, he does not imagine anything specific about the conditions of the ground. Finally, $[nec_{(p,\tau)}]r$ is true at any state, since all states in any cell of the τ-partition for p are states where there are plenty of bikes in Amsterdam, no matter whether r is on-topic relative to p or not.

With Definition 5.2.5 in place, we can now see that topic inclusion as well as the occurrence of imagination acts based on an input (and topic) can be expressed in $\mathcal{L}_{\mathsf{VI}}$. In fact, it is easy to check that the following hold for any VI model M, state w in M, $\varphi, \psi \in \mathcal{L}_{\mathsf{VI}}$, and $\tau \in T$:

1. $M, w \models [im_\varphi]\overline{\psi}$ iff $t(\psi) \sqsubseteq t(\varphi)$;

2. $M, w \models [im_{(\varphi,\tau)}]\overline{\psi}$ iff $t(\psi) \sqsubseteq t(\varphi)$;

[13] In particular, the operators are non-monotonic: $[im_\varphi]\psi$ can be true while $[im_{\varphi\wedge\chi}]\psi$ is false. Given *I jump the stream*, we imagine that we make it to the other side; but given *I jump the stream and I am carrying a 50kg backpack*, we imagine that we fall in the water.

3. $M, w \models \langle im_{\varphi} \rangle \overline{\psi}$ iff $f_{in}(\varphi, w) \neq \varnothing$;

4. $M, w \models \langle im_{(\varphi,\tau)} \rangle \overline{\psi}$ iff $f_{top}(\varphi, \tau, w) \neq \varnothing$.

Finally, let us select the class of models that are appropriate to represent ROMS, namely those where imagination acts are successful.

5.2.6. Definition (Appropriate VI model). A VI model M is said to be appropriate only if it satisfies the following condition (called *Success*): for any formula $\varphi \in \mathcal{L}_{\mathsf{VI}}$ and states v, w in M, if $v \in f_{in}(\varphi, w)$, then $M, v \models \varphi$.

Success has it that all closest φ-states are φ-states. It ensures that φ is true of any imagined scenario based on input φ, no matter under which topic φ is specified. Thus, given input φ, one always imagines that φ. (We mention that this may be controversial if imaginative resistance is taken seriously [Gendler, 2000].)

5.2.4 Axiomatization, Soundness, and Completeness

The axiom system VI is defined by the axioms and rules in Table 5.1. The six items at the top and the inclusion axioms are standard. Axioms in group (II) state the properties of the topic-driven imagination modalities. The first two axioms reveal that topic-driven imaginings obey two basic closure principles. According to $\mathsf{K}_{[im_{(\varphi,\tau)}]}$, when we imagine that an implication and its antecedent are true, we also imagine that the consequent is true. According to $\mathsf{C}_{[im_{(\varphi,\tau)}]}$ when we imagine that two propositions are true, we also imagine that their conjunction is true. Axioms $\mathsf{Ax1}$-$\mathsf{Ax3}$ concern topicality. $\mathsf{Ax1}$ expresses the relevance constraint proper of imagination acts: it says that ψ is imagined given input φ and topic τ only if it is on topic relative to φ. $\mathsf{Ax2}$ reflects the assumption that a sentence is about what its atoms are about: it says that every sentence consisting of propositional variables occurring in φ is on-topic relative to φ. $\mathsf{Ax3}$ simply states that topic inclusion is a transitive relation.

The bridge axioms in group (III) characterize the relation between the topic-driven imagination modalities and the other modalities of $\mathcal{L}_{\mathsf{VI}}$. $\mathsf{Ax4}$ reflects the fact that the topic of every sentence is fixed across all possible state, so that, if ψ is on-topic relative to φ at a state, it is on-topic relative to φ at all states. $\mathsf{Ax5}$ is a restricted principle of closure under strict implication: it states that, if a proposition is strictly implied by the input, then we imagine that that proposition is true *provided that* it is on-topic relative to the input. $\mathsf{Ax6}$ corresponds to the condition of no parting of indistinguishable states – no answer the agent can give to the questions determined by an input and topic can take apart states that give the same answers to those questions.

Finally, $\mathsf{Def}_{[im_{\varphi}]}$ reveals that $[im_{\varphi}]$ is definable in terms of $\Box$ and $[im_{(\varphi,\tau)}]$: we imagine that ψ is true given input φ precisely when we imagine that ψ is true, no matter which topic we choose and how we decide to specify it. As the proof of

(CPL)	Classical propositional tautologies	(MP)	From φ and $\varphi \to \psi$, infer ψ
($\mathsf{S5_A}$)	The axiom schemas of S5 for A	($\mathsf{RN_A}$)	From φ, infer $\mathsf{A}\varphi$
($\mathsf{S5}_\Box$)	The axiom schemas of S5 for $\Box$		
($\mathsf{S5}_{[nec_{(\varphi,\tau)}]}$)	The axiom schemas of S5 for $[nec_{(\varphi,\tau)}]$		
	(I) Inclusion axioms		
(Inc1)	$\mathsf{A}\psi \to \Box\psi$	(Inc2)	$\mathsf{A}\psi \to [\sim_{(\varphi,\tau)}]\psi$
	(II) Axioms for $[im_{(\varphi,\tau)}]$		
($\mathsf{K}_{[im_{(\varphi,\tau)}]}$)	$[im_{(\varphi,\tau)}](\psi \to \chi) \to ([im_{(\varphi,\tau)}]\psi \to [im_{(\varphi,\tau)}]\chi)$		
($\mathsf{C}_{[im_{(\varphi,\tau)}]}$)	$[im_{(\varphi,\tau)}]\psi \wedge [im_{(\varphi,\tau)}]\chi \to [im_{(\varphi,\tau)}](\psi \wedge \chi)$		
(Ax1)	$[im_{(\varphi,\tau)}]\psi \to [im_{(\varphi,\tau)}]\overline{\psi}$		
(Ax2)	$[im_{(\varphi,\tau)}]\overline{\psi}$, provided that $Var(\psi) \subseteq Var(\varphi)$		
(Ax3)	$[im_{(\varphi,\tau)}]\overline{\psi} \wedge [im_{(\psi,\tau)}]\overline{\chi} \to [im_{(\varphi,\tau)}]\overline{\chi}$		
	(III) Other bridge axioms		
(Ax4)	$[im_{(\varphi,\tau)}]\overline{\psi} \to \mathsf{A}[im_{(\varphi,\tau)}]\overline{\psi}$		
(Ax5)	$\mathsf{A}(\varphi \to \psi) \wedge [im_{(\varphi,\tau)}]\overline{\psi} \to [im_{(\varphi,\tau)}]\psi$		
(Ax6)	$[im_{(\varphi,\tau)}]\psi \to [im_{(\varphi,\tau)}][\sim_{(\varphi,\tau)}]\psi$		
($\mathsf{Def}_{[im_\varphi]}$)	$[im_\varphi]\psi \leftrightarrow \Box[im_{(\varphi,\tau)}]\psi$		

Table 5.1: The axiom system VI

soundness [Theorem 5.2.8] will make clear, $\mathsf{Def}_{[im_\varphi]}$ corresponds to the fact that the conditions on VI models ensure that, for any input φ, topic τ, and state w, $f_{in}(\varphi, w)$ is the union, for $w' \in R_\Box(w)$, of $f_{top}(\varphi, \tau, w')$. We decided to introduce f_{in} and f_{top} separately and then to include $\mathsf{Def}_{[im_\varphi]}$ as an axiom to highlight the role of the process of integration of the input (independently from topic selection) modeled by f_{in}.

$\mathsf{Def}_{[im_\varphi]}$ is key to prove the following proposition, which shows that the logic of $[im_\varphi]$ is analogous to the logic of $[im_{(\varphi,\tau)}]$.

5.2.7. Proposition. *The following are theorems of* VI:

($\mathsf{K}_{[im_\varphi]}$) $[im_\varphi](\psi \to \chi) \to ([im_\varphi]\psi \to [im_\varphi]\chi)$

($\mathsf{C}_{[im_\varphi]}$) $[im_\varphi]\psi \wedge [im_\varphi]\chi \to [im_\varphi](\psi \wedge \chi)$

(Thm1) $[im_\varphi]\psi \to [im_\varphi]\overline{\psi}$

(Thm2) $[im_\varphi]\overline{\psi}$, *provided that* $Var(\psi) \subseteq Var(\varphi)$

(Thm3) $[im_\varphi]\overline{\psi} \wedge [im_\psi]\overline{\chi} \rightarrow [im_\varphi]\overline{\chi}$

(Thm4) $\mathsf{A}(\varphi \rightarrow \psi) \wedge [im_\varphi]\overline{\psi} \rightarrow [im_\varphi]\psi$

(Thm5) $[im_\varphi]\psi \rightarrow [im_{(\varphi,\tau)}]\psi$

Proof:
Straightforward given axiom $\mathsf{Def}_{[im_\varphi]}$ and the axioms on operators $\Box$ and $[im_{(\varphi,\tau)}]$. $\Box$

5.2.8. Theorem. *The axiom system* VI *is sound with respect to the class of all appropriate* VI *models.*

We only present the details of the proof of the most interesting cases.

Proof:
Let M be an appropriate VI model and w a state in W. Checking that the axioms of S5 for A, $\Box$, and $[nec_{(\varphi,\tau)}]$ are valid and the rules MP and $\mathsf{RN_A}$ preserve validity is standard (recall that A is interpreted as a global modality and that $\Box$ and $[nec_{(\varphi,\tau)}]$ are interpreted in terms of equivalence relations between possible states). The validity of the two inclusion axioms depends on the fact that A is the global modality. Axioms $\mathsf{K}_{[im_{(\varphi,\tau)}]}$ and $\mathsf{C}_{[im_{(\varphi,\tau)}]}$ are valid because of the semantic clause for $[im_{(\varphi,\tau)}]$ and the def. of t. We only spell out the proof that $\mathsf{C}_{[im_{(\varphi,\tau)}]}$ is valid.

($\mathsf{C}_{[im_{(\varphi,\tau)}]}$) Suppose that $M, w \models [im_{(\varphi,\tau)}]\psi \wedge [im_{(\varphi,\tau)}]\chi$. By Def. 5.2.5, (1) $f_{top}(\varphi, \tau, w) \subseteq [\![\psi]\!]$, (2) $t(\psi) \sqsubseteq t(\varphi)$, (3) $f_{top}(\varphi, \tau, w) \subseteq [\![\chi]\!]$, (4) $t(\chi) \sqsubseteq t(\varphi)$. From (1) and (3) it follows that (5) $f_{top}(\varphi, \tau, w) \subseteq [\![\psi \wedge \chi]\!]$; from (2) and (4) it follows that $t(\psi) \oplus t(\chi) \sqsubseteq t(\varphi)$, and so (6) $t(\psi \wedge \chi) \sqsubseteq t(\varphi)$ by the def. of t. From (5), (6), and Def. 5.2.5 we get $M, w \models [im_{(\varphi,\tau)}](\psi \wedge \chi)$.

The validity of Ax1 and Ax2 is an immediate consequence of item 1 on p. 125 and the def. of t; the validity of Ax3 depends, in addition, on the transitivity of $\sqsubseteq$. Axiom Ax4 is valid because the topic of every formula is invariant across possible worlds. The validity of the remaining axioms depends on the properties of appropriate VI models.

(Ax5) Suppose that $M, w \models \mathsf{A}(\varphi \rightarrow \psi) \wedge [im_{(\varphi,\tau)}]\overline{\psi}$. Then, by Def. 5.2.5 and item 1 on p. 125, (1) $[\![\varphi]\!] \subseteq [\![\psi]\!]$ and (2) $t(\psi) \sqsubseteq t(\varphi)$. By the condition of success, $f_{in}(\varphi, w) \subseteq [\![\varphi]\!]$ and, by the condition of specification of the basic initial scenario, $f_{top}(\varphi, \tau, w) \subseteq f_{in}(\varphi, w)$. Hence, (3) $f_{top}(\varphi, \tau, w) \subseteq [\![\varphi]\!]$. From (1) and (3) we get (4) $f_{top}(\varphi, \tau, w) \subseteq [\![\psi]\!]$. Given Def. 5.2.5, (2) and (4) suffice to conclude $M, w \models [im_{(\varphi,\tau)}]\psi$.

(Ax6) Suppose that $M, w \models [im_{(\varphi,\tau)}]\psi$. By Def. 5.2.5, (1) $f_{top}(\varphi, \tau, w) \subseteq [\![\psi]\!]$ and (2) $t(\psi) \sqsubseteq t(\varphi)$. Take any world $v \in f_{top}(\varphi, \tau, w)$. By the condition of no parting of

indistinguishable worlds, for all v' s.t. $v \sim_{(\varphi,\tau)} v'$, $v' \in f_{top}(\varphi, \tau, w)$, and so $v' \in [\![\psi]\!]$ by (1). But then (3) $f_{top}(\varphi, \tau, w) \subseteq [\![[nec_{(\varphi,\tau)}]\psi]\!]$ by Def. 5.2.5. In addition, since $t([nec_{(\varphi,\tau)}]\psi) = t(\varphi) \oplus t(\psi)$ by the def. of t, (2) implies that $t([nec_{(\varphi,\tau)}]\psi) = t(\varphi)$ and thus (4) $t([nec_{(\varphi,\tau)}]\psi) \sqsubseteq t(\varphi)$. Given Def. 5.2.5, (3) and (4) suffice to conclude $M, w \models [im_{(\varphi,\tau)}][nec_{(\varphi,\tau)}]\psi$.

($\mathsf{Def}_{[im_\varphi]}$) Observe that, by Def. 5.2.5, we have that:

1. $M, w \models [im_\varphi]\psi$ iff $f_{in}(\varphi, w) \subseteq [\![\psi]\!]$ and $t(\psi) \sqsubseteq t(\varphi)$;
2. $M, w \models \Box[im_{(\varphi,\tau)}]\psi$ iff $\bigcup_{w' \in R_\Box(w)} f_{top}(\varphi, \tau, w') \subseteq [\![\psi]\!]$ and $t(\psi) \sqsubseteq t(\varphi)$.

Hence, it suffices to prove that $f_{in}(\varphi, w) = \bigcup_{w' \in R_\Box(w)} f_{top}(\varphi, \tau, w')$. Suppose, first, that $v \in f_{in}(\varphi, w)$. Then, by the condition of ability to select any answer, there is $w' \in R_\Box(w)$ s.t. $v \in f_{top}(\varphi, \tau, w')$. Thus, $v \in \bigcup_{w' \in R_\Box(w)} f_{top}(\varphi, \tau, w')$. For the other direction, suppose that $v \in \bigcup_{w' \in R_\Box(w)} f_{top}(\varphi, \tau, w')$, so that there is a state w' s.t. (1) $w' \in R_\Box(w)$ and (2) $v \in f_{top}(\varphi, \tau, w')$. From (2) and the condition of specification of the basic initial scenario, it follows that $v \in f_{in}(\varphi, w')$. In addition, by the condition of no choice of the basic initial scenario, (1) implies that $f_{in}(\varphi, w') = f_{in}(\varphi, w)$, whence the result. □

5.2.9. Theorem. *The axiom system* VI *is strongly complete with respect to the class of all appropriate* VI *models.*

The proof of Theorem 5.2.9 proceeds via the construction of a canonical model for VI, which, in turn, is based on the construction of a canonical topic model for VI. The construction is inspired by the one presented by Giordani [2019]. In the following, we will assume the usual definitions of VI-consistent set of formulas and maximally VI-consistent set of formulas (for short: *mcs*). In addition, standard results concerning the properties of mcs as well as Lindenbaum's Lemma will be repeatedly used in the proofs below without explicit mention.[14] For any set of formulas $\Gamma \subseteq \mathcal{L}_{\mathsf{VI}}$ and $\varphi \in \mathcal{L}_{\mathsf{VI}}$, we write $\vdash_{\mathsf{VI}} \varphi$ if φ is a theorem of $\mathcal{L}_{\mathsf{VI}}$ and $\Gamma \vdash_{\mathsf{VI}} \varphi$ if φ is deducible in VI from Γ.

Let us start from the canonical topic model for VI. The key idea is to define an equivalence relation over $\mathcal{L}_{\mathsf{VI}}$ and identify topics with equivalence classes determined by this relation. More specifically, let $\mathcal{W}$ be the set of all mcs of VI. For any $w \in \mathcal{W}$, define $\sim_w \subseteq \mathcal{L}_{\mathsf{VI}} \times \mathcal{L}_{\mathsf{VI}}$ by setting: for all $\varphi, \psi \in \mathcal{L}_{\mathsf{VI}}$,

$$\varphi \sim_w \psi \text{ iff } [im_\varphi]\overline{\psi} \wedge [im_\psi]\overline{\varphi} \in w$$

[14] The definitions of (maximally) consistent set of formulas and the statements of the relevant lemmas can be found in Appendix A.1, on page 203. More details and proofs can be found in Blackburn et al. [2001, Chapter 4.2].

5.2.10. Lemma. *For all $w \in \mathcal{W}$, $\sim_w$ is an equivalence relation on $\mathcal{L}_{\mathsf{VI}}$.*

Proof:
Let $w \in \mathcal{W}$. We check that $\sim_w$ is (1) reflexive, (2) transitive, and (3) symmetric.

1. Take any $\varphi \in \mathcal{L}_{\mathsf{VI}}$. By Thm2, $[im_\varphi]\overline{\varphi} \in w$, hence $\varphi \sim_w \varphi$ by the def. of $\sim_w$.
2. Take any $\varphi, \psi, \chi \in \mathcal{L}_{\mathsf{VI}}$ s.t. $\varphi \sim_w \psi$ and $\psi \sim_w \chi$. By the def. of $\sim_w$, $[im_\varphi]\overline{\psi} \wedge [im_\psi]\overline{\varphi} \in w$ and $[im_\psi]\overline{\chi} \wedge [im_\chi]\overline{\psi} \in w$. So, $[im_\varphi]\overline{\psi} \wedge [im_\psi]\overline{\chi} \in w$ and $[im_\chi]\overline{\psi} \wedge [im_\psi]\overline{\varphi} \in w$. By Thm3, $[im_\varphi]\overline{\chi} \wedge [im_\chi]\overline{\psi} \in w$. Hence, $\varphi \sim_w \chi$ by the def. of $\sim_w$.
3. Take any $\varphi, \psi \in \mathcal{L}_{\mathsf{VI}}$ s.t. $\varphi \sim_w \psi$. By the def. of $\sim_w$, $[im_\varphi]\overline{\psi} \wedge [im_\psi]\overline{\varphi} \in w$. So, $[im_\psi]\overline{\varphi} \wedge [im_\varphi]\overline{\psi} \in w$. That is, $\psi \sim_w \varphi$ by the def. of $\sim_w$.

□

For any $w \in \mathcal{W}$, we write $\mathcal{L}_{\mathsf{VI}} / \sim_w$ for the quotient set of $\mathcal{L}_{\mathsf{VI}}$ by $\sim_w$ and $[\varphi]_w$ for the equivalence class of φ in $\mathcal{L}_{\mathsf{VI}} / \sim_w$.

5.2.11. Definition (Canonical topic VI model for w_0). Let w_0 be a mcs. The canonical topic VI model for w_0 is the tuple $\mathcal{T}^c = \langle T^c, \oplus^c, t^c \rangle$, where

- $T^c = \mathcal{L}_{\mathsf{VI}} / \sim_{w_0}$;
- $\oplus^c : T^c \times T^c \to T^c$ is such that, for all $\varphi, \psi \in \mathcal{L}_{\mathsf{VI}}$, $[\varphi]_{w_0} \oplus^c [\psi]_{w_0} = [\varphi \wedge \psi]_{w_0}$;
- $t^c : Prop \to T^c$ is such that, for all $p \in Prop$, $t^c(p) = [p]_{w_0}$.

The topic assignment t^c is extended to the whole of $\mathcal{L}_{\mathsf{VI}}$ as in Definition 5.2.3: for any $\varphi \in \mathcal{L}_{\mathsf{VI}}$, if $Var(\varphi) = \{p_1, \ldots, p_n\}$, then $t^c(\varphi) = t^c(p_1) \oplus^c \cdots \oplus^c t^c(p_n)$.

5.2.12. Lemma. *The canonical topic VI model $\mathcal{T}^c$ is a topic VI model.*

Proof:
Let us start by checking that $\oplus^c$ is well defined. Suppose that (1) $\varphi_1 \sim_{w_0} \varphi_2$ and (2) $\psi_1 \sim_{w_0} \psi_2$. We need to prove that $(\varphi_1 \wedge \psi_1) \sim_{w_0} (\varphi_2 \wedge \psi_2)$. By the def. of $\sim_{w_0}$, (1) implies that (3) $[im_{\varphi_1}]\overline{\varphi_2} \wedge [im_{\varphi_2}]\overline{\varphi_1} \in w_0$ and (2) implies that (4) $[im_{\psi_1}]\overline{\psi_2} \wedge [im_{\psi_2}]\overline{\psi_1} \in w_0$. By Thm2, (5) $[im_{\varphi_1 \wedge \psi_1}]\overline{\varphi_1} \in w_0$ and (6) $[im_{\varphi_1 \wedge \psi_1}]\overline{\psi_1} \in w_0$. From (3), (5), and Thm3 it follows that (7) $[im_{\varphi_1 \wedge \psi_1}]\overline{\varphi_2} \in w_0$. Similarly from (4), (6), and Thm3 it follows that (8) $[im_{\varphi_1 \wedge \psi_1}]\overline{\psi_2}$. (7) and (8) imply that $[im_{\varphi_1 \wedge \psi_1}](\overline{\varphi_2} \wedge \overline{\psi_2}) \in w_0$ by $\mathsf{C}_{[im_\varphi]}$, which, in turn, implies that $[im_{\varphi_1 \wedge \psi_1}]\overline{\varphi_2 \wedge \psi_2} \in w_0$ by the def. of $\overline{\varphi}$. By reasoning in an analogous way, we establish that $[im_{\varphi_2 \wedge \psi_2}]\overline{\varphi_1 \wedge \psi_1} \in w_0$, and so $(\varphi_1 \wedge \psi_1) \sim_{w_0} (\varphi_2 \wedge \psi_2)$ by the def. of $\sim_{w_0}$. It remains to check that $\oplus^c$ is (1) idempotent, (2) commutative, (3) associative, and (4) that the topic assignment t^c is extended to the whole $\mathcal{L}_{\mathsf{VI}}$ in a consistent way.

1. For any $\varphi \in \mathcal{L}_{\mathsf{VI}}$, $[im_{\varphi\wedge\varphi}]\overline{\varphi} \wedge [im_{\varphi}]\overline{\varphi\wedge\varphi} \in w_0$ by Thm2, and so $[\varphi]_{w_0} \oplus^c [\varphi]_{w_0} = [\varphi\wedge\varphi]_{w_0} = [\varphi]_{w_0}$ by the def. of $\oplus^c$ and $\sim_w$.

2. For any $\varphi, \psi \in \mathcal{L}_{\mathsf{VI}}$, $[im_{\varphi\wedge\psi}]\overline{\psi\wedge\varphi} \wedge [im_{\psi\wedge\varphi}]\overline{\varphi\wedge\psi} \in w_0$ by Thm2, and so $[\varphi]_{w_0} \oplus^c [\psi]_{w_0} = [\varphi\wedge\psi]_{w_0} = [\psi\wedge\varphi]_{w_0} = [\psi]_{w_0} \oplus^c [\varphi]_{w_0}$ by the def. of $\oplus^c$ and $\sim_w$.

3. By reasoning as in 2, it is easy to prove that, for any $\varphi, \psi, \chi \in \mathcal{L}_{\mathsf{VI}}$, $[\varphi\wedge(\psi\wedge\chi)]_{w_0} = [(\varphi\wedge\psi)\wedge\chi]_{w_0}$, and so $[\varphi]_{w_0} \oplus^c ([\psi]_{w_0} \oplus^c [\chi]_{w_0}) = ([\varphi]_{w_0} \oplus^c [\psi]_{w_0}) \oplus^c [\chi]_{w_0}$.

4. Thm2 ensures that, for all $\varphi \in \mathcal{L}_{\mathsf{VI}}$, if $Var(\varphi) = \{p_1, \ldots, p_n\}$, then $(p_1 \wedge \cdots \wedge p_n) \sim_w \varphi$. We then have:
$t^c(\varphi) = t^c(p_1) \oplus^c \cdots \oplus^c t^c(p_n) = [p_1]_{w_0} \oplus^c \cdots \oplus^c [p_n]_{w_0} = [p_1 \wedge \cdots \wedge p_n]_{w_0} = [\varphi]_{w_0}$

□

The relation $\sqsubseteq^c \subseteq T^c \times T^c$ of topic inclusion is defined as in Section 5.2.2.

5.2.13. Remark. It is easy to see that, for all $\varphi, \psi \in \mathcal{L}_{\mathsf{VI}}$, $t^c(\psi) \sqsubseteq t^c(\varphi)$ iff $[im_\varphi]\overline{\psi} \in w_0$. To be sure:

(a)	$t^c(\psi) \sqsubseteq^c t^c(\varphi)$	$\Leftrightarrow$	$t^c(\psi) \oplus^c t^c(\varphi) = t^c(\varphi)$	by the def. of $\sqsubseteq^c$
		$\Leftrightarrow$	$[\psi\wedge\varphi]_{w_0} = [\varphi]_{w_0}$	by the def. of t^c
		$\Rightarrow$	$[im_\varphi]\overline{\psi\wedge\varphi} \in w_0$	by the def. of $\sim_{w_0}$
		$\Rightarrow$	$[im_{\psi\wedge\varphi}]\overline{\psi} \in w_0$	by Thm2
		$\Rightarrow$	$[im_\varphi]\overline{\psi} \in w_0$	by Thm3
(b)	$[im_\varphi]\overline{\psi} \in w_0$	$\Rightarrow$	$[im_\varphi]\overline{\psi} \wedge [im_\varphi]\overline{\varphi} \in w_0$	by Thm2
		$\Rightarrow$	$[im_\varphi]\overline{\psi\wedge\varphi} \in w_0$	by $\mathsf{C}_{[im_\varphi]}$ and def. $\overline{\varphi}$
		$\Rightarrow$	$[im_\varphi]\overline{\psi\wedge\varphi} \wedge [im_{\psi\wedge\varphi}]\overline{\varphi} \in w_0$	by Thm2
		$\Leftrightarrow$	$[\psi\wedge\varphi]_{w_0} = [\varphi]_{w_0}$	by the def. of t^c
		$\Leftrightarrow$	$t^c(\psi) \oplus^c t^c(\varphi) = t^c(\varphi)$	by the def. of $\oplus^c$
		$\Leftrightarrow$	$t^c(\psi) \sqsubseteq^c t^c(\varphi)$	by the def. of $\sqsubseteq^c$

Before we define the canonical VI model, let us introduce a final bit of notation. Where w is a mcs and $\blacksquare \in \{\mathsf{A}, \Box\}$, $\varphi \in \mathcal{L}_{\mathsf{VI}}$, and $\tau \in T$, let

1. $w/\blacksquare = \{\varphi \in \mathcal{L}_{\mathsf{VI}} \mid \blacksquare\varphi \in w\}$;

2. $w /\!\!/ [im_{(\varphi,\tau)}] = \{\chi \in \mathcal{L}_{\mathsf{VI}} \mid \mathsf{A}(\chi \to \psi) \wedge [im_{(\varphi,\tau)}]\chi \in w, \text{for some } \psi \in \mathcal{L}_{\mathsf{VI}}\}$

5.2.14. Remark. It follows immediately from the logic of A and from axiom $\mathsf{C}_{[im_{(\varphi,\tau)}]}$ that $w /\!\!/ [im_{(\varphi,\tau)}]$ is closed under finite conjunction. That is, for any finite set $\Gamma \subseteq \mathcal{L}_{\mathsf{VI}}$, if $\Gamma \subseteq w /\!\!/ [im_{(\varphi,\tau)}]$, then $\bigwedge\Gamma \in w /\!\!/ [im_{(\varphi,\tau)}]$.

5.2.15. Definition (Canonical VI model for w_0). Let w_0 be a mcs. The canonical VI model for w_0 is a tuple $M^c = \langle \mathcal{T}^c, W^c, R^c_\Box, f^c_{in}, f^c_{top}, \sim^c, \nu^c \rangle$, where

- $\mathcal{T}^c = \langle T^c, \oplus^c, t^c \rangle$ is the canonical VI topic model for w_0;
- $W^c = \{w \in \mathcal{W} \mid w_0/\mathsf{A} \subseteq w\}$;
- $R^c_\Box \subseteq W^c \times W^c$ is such that $wR^c_\Box w'$ iff $w/\Box \subseteq w'$;
- $f^c_{in} : \mathcal{L}_{\mathsf{VI}} \times W^c \to 2^{W^c}$ is such that $v \in f^c_{in}(\varphi, w)$ iff, for some $w' \in W^c$ and $\tau \in T$, $wR^c_\Box w'$ and $v \in f^c_{top}(\varphi, \tau, w')$;
- $f^c_{top} : \mathcal{L}_{\mathsf{VI}} \times T \times W^c \to 2^{W^c}$ is such that $v \in f^c_{top}(\varphi, \tau, w)$ iff $w /\!\!/ [im_{(\varphi,\tau)}] \subseteq v$;
- $\sim^c$: $\mathcal{L}_{\mathsf{VI}} \times T \to 2^{W^c \times W^c}$ is such that $w \sim^c_{(\varphi,\tau)} v$ iff $w/[nec_{(\varphi,\tau)}] \subseteq v$;
- $\nu^c : Prop \to 2^{W^c}$ is such that $w \in \nu^c(p)$ iff $p \in w$.

The next three lemmas are crucial to prove the Truth Lemma [Lemma 5.2.20].

5.2.16. Lemma. *For all $\varphi \in \mathcal{L}_{\mathsf{VI}}$, $\varphi \in w_0/\mathsf{A}$ iff, for all $w \in W^c$, $\varphi \in w$.*

Proof:
Take any $\varphi \in \mathcal{L}_{\mathsf{VI}}$. If $\varphi \in w_0/\mathsf{A}$, then, for all $w \in W^c$, $\varphi \in w$ by the def. of W^c. For the other direction, suppose that, for all $w \in W^c$, $\varphi \in W^c$. By the def. of W^c, this means that, for all $w \in \mathcal{W}$ s.t. $w_0/\mathsf{A} \subseteq w$, $\varphi \in w$. By an immediate corollary of Lindenbaum's Lemma, it follows that $w_0/\mathsf{A} \vdash_{\mathsf{VI}} \varphi$. That is, there is a finite set $\Gamma \subseteq w_0/\mathsf{A}$ s.t. $\vdash_{\mathsf{VI}} \bigwedge \Gamma \to \varphi$. By the logic of A, $\vdash_{\mathsf{VI}} \bigwedge_{\psi \in \Gamma} \mathsf{A}\psi \to \mathsf{A}\varphi$. Since $\Gamma \subseteq w_0/\mathsf{A}$, $\bigwedge_{\psi \in \Gamma} \mathsf{A}\psi \in w_0$ by the def. of w_0/A, and so $\mathsf{A}\varphi \in w_0$. Hence, $\varphi \in w_0/\mathsf{A}$ by the def. of w_0/A. □

5.2.17. Lemma (Existence Lemma). *For all $w \in W^c$ and $\varphi, \psi \in \mathcal{L}_{\mathsf{VI}}$, (a) if $\Diamond\varphi \in w$, then there is $w' \in W^c$ s.t. $w/\Box \subseteq w'$ and $\varphi \in w'$, and (b) if $\langle nec_{(\varphi,\tau)} \rangle \psi \in w$, then there is $w' \in W^c$ s.t. $w/[nec_{(\varphi,\tau)}] \subseteq w'$ and $\varphi \in w'$.*

Proof:
We only prove (a) (the proof of (b) is analogous). Take any $w \in W^c$ and $\varphi \in \mathcal{L}_{\mathsf{VI}}$ such that $\Diamond\varphi \in w$. Suppose, toward contradiction, that the set $w_0/\mathsf{A} \cup w/\Box \cup \{\varphi\}$ is inconsistent. Then, there are finite sets $\Gamma \subseteq w_0/\mathsf{A}$ and $\Delta \subseteq w/\Box$ s.t. $\vdash_{\mathsf{VI}} \bigwedge \Gamma \wedge \bigwedge \Delta \to \neg\varphi$. By the logic of $\Box$, (1) $\vdash_{\mathsf{VI}} \bigwedge_{\psi \in \Gamma} \Box\psi \wedge \bigwedge_{\chi \in \Delta} \Box\chi \to \Box\neg\varphi$. Since $\Delta \subseteq w/\Box$, (2) $\bigwedge_{\chi \in \Delta} \Box\chi \in w$. In addition, since $\Gamma \subseteq w_0/\mathsf{A}$, for all $\psi \in \Gamma$, $\mathsf{A}\psi \in w_0$ and so $\mathsf{AA}\psi \in w_0$ by $\mathsf{S5_A}$. By the def. of W^c, for all $\psi \in \Gamma$, $\mathsf{A}\psi \in w$ and so $\Box\psi \in w$ by axiom $\mathsf{Inc1}$. Hence, (3) $\bigwedge_{\psi \in \Gamma} \Box\psi \in w$. It follows from (1), (2), and (3) that $\Box\neg\varphi \in w$, so that $\Diamond\varphi \notin w$, against the hypothesis. Therefore, $w_0/\mathsf{A} \cup w/\Box \cup \{\varphi\}$ is consistent and can be extended to a mcs w'. Since $w_0/\mathsf{A} \subseteq w'$, $w' \in W^c$. In addition, $w/\Box \subseteq w'$ and $\varphi \in w'$. □

The proof of Lemma 5.2.19 below relies on the following proposition.

5.2.18. PROPOSITION. *The following are theorems of* VI*:*

(Thm6) $[im_{(\varphi,\tau)}]\psi \to [im_\varphi]\overline{\psi}$
(Thm7) $\mathsf{A}(\chi \to \psi) \wedge [im_\varphi]\chi \wedge [im_\varphi]\overline{\psi} \to [im_\varphi]\psi$

Proof:
The proof of Thm6 is as follows:

(1)	$\vdash_{\mathsf{VI}} [im_{(\varphi,\tau)}]\psi \to [im_{(\varphi,\tau)}]\overline{\psi}$	by Ax1
(2)	$\vdash_{\mathsf{VI}} [im_{(\varphi,\tau)}]\overline{\psi} \to \mathsf{A}[im_{(\varphi,\tau)}]\overline{\psi}$	by Ax4
(3)	$\vdash_{\mathsf{VI}} \mathsf{A}[im_{(\varphi,\tau)}]\overline{\psi} \to [\sim_{(\varphi,\tau)}][im_{(\varphi,\tau)}]\overline{\psi}$	by Inc1
(4)	$\vdash_{\mathsf{VI}} [\sim_{(\varphi,\tau)}][im_{(\varphi,\tau)}]\overline{\psi} \to [im_\varphi]\overline{\psi}$	by $\mathsf{Def}_{[im_\varphi]}$
(5)	$\vdash_{\mathsf{VI}} [im_{(\varphi,\tau)}]\psi \to [im_\varphi]\overline{\psi}$	from (1)-(4)

The proof of Thm7 is as follows:

(1)	$\vdash_{\mathsf{VI}} [im_\varphi]\chi \wedge [im_\varphi]\overline{\psi} \to [im_\varphi]\overline{\chi} \wedge [im_\varphi]\overline{\psi}$	by Thm1
(2)	$\vdash_{\mathsf{VI}} [im_\varphi]\overline{\chi} \wedge [im_\varphi]\overline{\psi} \to [im_\varphi](\overline{\chi} \wedge \overline{\psi})$	by $\mathsf{C}_{[im_\varphi]}$
(3)	$\vdash_{\mathsf{VI}} [im_\varphi](\overline{\chi} \wedge \overline{\psi}) \to [im_\varphi](\overline{\chi \to \psi})$	by the def. of $\overline{\varphi}$
(4)	$\vdash_{\mathsf{VI}} \mathsf{A}(\varphi \to (\chi \to \psi)) \wedge [im_\varphi](\overline{\chi \to \psi}) \to [im_\varphi](\chi \to \psi)$	by Thm4
(5)	$\vdash_{\mathsf{VI}} \mathsf{A}(\chi \to \psi) \to \mathsf{A}(\varphi \to (\chi \to \psi))$	by $\mathsf{S5_A}$
(6)	$\vdash_{\mathsf{VI}} \mathsf{A}(\chi \to \psi) \wedge [im_\varphi](\overline{\chi \to \psi}) \to [im_\varphi](\chi \to \psi)$	from (4)-(5)
(7)	$\vdash_{\mathsf{VI}} \mathsf{A}(\chi \to \psi) \wedge [im_\varphi]\chi \wedge [im_\varphi]\overline{\psi} \to [im_\varphi](\chi \to \psi)$	from (1)-(3), (6)
(8)	$\vdash_{\mathsf{VI}} [im_\varphi]\chi \wedge [im_\varphi](\chi \to \psi) \to [im_\varphi]\psi$	by $\mathsf{K}_{[im_\varphi]}$
(9)	$\vdash_{\mathsf{VI}} \mathsf{A}(\chi \to \psi) \wedge [im_\varphi]\chi \wedge [im_\varphi]\overline{\psi} \to [im_\varphi]\psi$	from (7)-(8)

□

5.2.19. LEMMA. *For all $w \in W^c$, $\varphi, \psi \in \mathcal{L}_{\mathsf{VI}}$, and $\tau \in T$,*

$[im_{(\varphi,\tau)}]\psi \in w$ *iff* $[im_\varphi]\overline{\psi} \in w$ *and, for all v s.t. $w /\!\!/ [im_{(\varphi,\tau)}] \subseteq v$, $\varphi \in v$.*

Proof:
First, suppose that (1) $[im_{(\varphi,\tau)}]\psi \in w$. By Thm6, (2) $[im_\varphi]\overline{\psi} \in w$. In addition, by $\mathsf{RN_A}$ and (1), $\mathsf{A}(\psi \to \psi) \wedge [im_{(\varphi,\tau)}]\psi \in w$. Hence, by the def. of $w /\!\!/ [im_{(\varphi,\tau)}]$, $\psi \in v$ for all v s.t. $w /\!\!/ [im_{(\varphi,\tau)}] \subseteq v$. For the other direction, suppose that (1′) $[im_\varphi]\overline{\psi} \in w$ and (2′) $\varphi \in v$, for all $v \in W^c$ s.t. $w /\!\!/ [im_{(\varphi,\tau)}] \subseteq v$. By an immediate corollary of Lindenbaum's Lemma, (2′) implies that $w_0/\mathsf{A} \cup w /\!\!/ [im_{(\varphi,\tau)}] \vdash_{\mathsf{VI}} \varphi$. Hence, there are finite sets $\Delta_1 \subseteq w_0/\mathsf{A}$ and $\Delta_2 \subseteq w /\!\!/ [im_{(\varphi,\tau)}]$ such that $\vdash_{\mathsf{VI}} \bigwedge \Delta_1 \wedge \bigwedge \Delta_2 \to \varphi$. By the logic of A, $\vdash_{\mathsf{VI}} \mathsf{A}(\bigwedge \Delta_1 \wedge \bigwedge \Delta_2 \to \varphi)$, and so (3′) $\mathsf{A}(\bigwedge \Delta_1 \wedge \bigwedge \Delta_2 \to \varphi) \in w$. Since $\Delta_1 \subseteq w_0/\mathsf{A}$, (4′) $\mathsf{A}(\bigwedge \Delta_1) \in w$ by the def. of w_0/A, W^c, and the logic of A. It

follows from (3′) and (4′) that (5′) $\mathsf{A}(\bigwedge \Delta_2 \to \varphi) \in w$. Now, since $\Delta_2 \subseteq w /\!\!/ [im_{(\varphi,\tau)}]$, $\bigwedge \Delta_2 \in w /\!\!/ [im_{(\varphi,\tau)}]$ by Remark 5.2.14. Hence, by the def. of $w /\!\!/ [im_{(\varphi,\tau)}]$, there is $\chi \in \mathcal{L}_{\mathsf{VI}}$ s.t. (6′) $\mathsf{A}(\chi \to \bigwedge \Delta_2) \wedge [im_\varphi]\chi \in w$. From (5′) and (6′), we obtain (7′) $\mathsf{A}(\chi \to \varphi) \wedge [im_\varphi]\chi \in w$. (1′), (7′), and Thm7 imply that $[im_\varphi]\psi \in w$. Thus, $[im_{(\varphi,\tau)}]\psi \in w$ by Thm5. □

5.2.20. Lemma (Truth Lemma). *For all $w \in W^c$ and $\varphi \in \mathcal{L}_{\mathsf{VI}}$,*

$$M^c, w \models \varphi \text{ iff } \varphi \in w$$

Proof:
The proof is by induction on the complexity of φ. The cases for the propositional variables, Booleans, $\varphi := \mathsf{A}\psi$, $\varphi := \Box\psi$, $\varphi := [nec_{(\varphi,\tau)}]$ are standard (the latter three cases use, resp., Lemmas 5.2.16 and 5.2.17). We prove the remaining cases.

1. $\varphi := [im_{(\psi,\tau)}]\chi$

$M^c, w \models [im_{(\psi,\tau)}]\chi$ iff for all $v \in W^c$ s.t. $v \in f^c_{top}(\psi, \tau, w), M^c, v \models \chi$, and $t^c(\chi) \sqsubseteq^c t^c(\psi)$ (Def. 5.2.5)

$M^c, w \models [im_{(\psi,\tau)}]\chi$ iff for all $v \in W^c$ s.t. $w /\!\!/ [im_{(\psi,\tau)}] \subseteq v, M^c, v \models \chi$, and $t^c(\chi) \sqsubseteq^c t^c(\psi)$ (by the def. of f^c_{top})

iff for all $v \in W^c$ s.t. $w /\!\!/ [im_{(\psi,\tau)}] \subseteq v, \chi \in w$, and $[im_\psi]\overline{\chi} \in w_0$ (by induction hypothesis and Rem. 5.2.13)

iff $[im_{(\psi,\tau)}]\chi \in w$ (by Lem. 5.2.19)

2. $\varphi := [im_\psi]\chi$

$M^c, w \models [im_\psi]\chi$ iff for all $v \in W^c$ s.t. $v \in f^c_{in}(\psi, w), M^c, v \models \chi$ and $t^c(\chi) \sqsubseteq^c t^c(\psi)$ (Def. 5.2.5)

$M^c, w \models [im_\psi]\chi$ iff for all $w', v \in W^c$ s.t. $wR^c_\Box w'$ and $v \in f^c_{top}(\psi, \tau, w'), M^c, v \models \chi$, and $t^c(\chi) \sqsubseteq^c t^c(\psi)$ (by the def. of f^c_{in})

iff for all $w', v \in W^c$ s.t. $wR^c_\Box w'$, $M^c, w' \models [im_{(\psi,\tau)}]\chi$, and $[im_\psi]\overline{\chi} \in w_0$ (Def. 5.2.5 and Rem. 5.2.13)

iff for all $w', v \in W^c$ s.t. $wR^c_\Box w'$, $[im_{(\psi,\tau)}]\chi \in w'$, and $[im_\psi]\overline{\chi} \in w_0$ (following the same steps as in case 1)

iff $\Box[im_{(\psi,\tau)}]\chi \in w$ and $[im_\psi]\overline{\chi} \in w$ (Lem. 5.2.17 and def. of $R^c_\Box$)

iff $[im_\psi]\chi \in w$ (by $\mathsf{Def}_{[im_\varphi]}$ and Thm1)

□

5.2.21. LEMMA. *The canonical* VI *model* M^c *is an appropriate* VI *model.*

Proof:
We have already proved that $\mathcal{T}^c$ is a topic VI model [see Lem. 5.2.12], so we focus on the other properties of appropriate VI models.

1. $R^c_\Box$ *and* $\sim^c_{(\varphi,\tau)}$ *are equivalence relations.*

   This follows from standard results in modal logic, as VI includes the axioms of S5 for $\Box$ and $[\sim_{(\varphi,\tau)}]$.

2. M^c *satisfies the condition of no choice of the basic initial scenario.*

   Let $w, v \in W^c$ be s.t. (1) $wR^c_\Box v$. Take any $u \in f^c_{in}(\varphi, w)$. By the def. of $f^c_{in}(\varphi, w)$, there is $w' \in W^c$ and $\tau \in T$ s.t. (2) $wR^c_\Box w'$ and (3) $u \in f^c_{top}(\varphi, \tau, w')$. Since $R^c_\Box$ is an equivalence relation, (1)-(3) imply that there is $w' \in W^c$ and $\tau \in T^c$ s.t. $vR^c_\Box w'$ and $u \in f^c_{top}(\varphi, \tau, w')$. By the def. of f^c_{in}, $u \in f^c_{in}(\varphi, v)$. We can then conclude that $f^c_{in}(\varphi, w) \subseteq f^c_{in}(\varphi, v)$. An analogous argument shows that $f^c_{in}(\varphi, v) \subseteq f^c_{in}(\varphi, w)$.

3. M^c *satisfies the condition of specification of the basic initial scenario.*

   Consider $w, v \in W^c$ s.t. $v \in f^c_{top}(\varphi, \tau, w)$. Since $R^c_\Box$ is reflexive, $wR^c_\Box w$ and $v \in f^c_{top}(\varphi, \tau, w)$, and so $v \in f^c_{in}(\varphi, w)$ by the def. of f^c_{in}.

4. M^c *satisfies the condition of no partying of indistinguishable worlds.*

   Consider $w, v \in W^c$ s.t. $v \in f^c_{top}(\varphi, \tau, w)$. Take any $u \in W^c$ s.t. $v \sim^c_{(\varphi,\tau)} u$. As $\sim^c_{(\varphi,\tau)}$ is an equivalence relation, $u \sim^c_{(\varphi,\tau)} v$. By the def. of $\sim^c$, $v/[\sim_{(\varphi,\tau)}] \subseteq u$. We want to show that $u \in f^c_{top}(\varphi, \tau, w)$. So, consider any $\psi, \chi \in \mathcal{L}_{\mathsf{VI}}$ s.t. $\mathsf{A}(\chi \to \psi) \wedge [im_{(\varphi,\tau)}]\chi \in w$. By the logic of A, the logic of $[nec_{(\varphi,\tau)}]$, and axiom Ax6, $\mathsf{A}([nec_{(\varphi,\tau)}]\psi \to [nec_{(\varphi,\tau)}]\chi) \wedge [im_{(\varphi,\tau)}][nec_{(\varphi,\tau)}]\chi \in w$. As $v \in f^c_{top}(\varphi, \tau, w)$, it follows that $[nec_{(\varphi,\tau)}]\chi \in v$ by the def. of f^c_{top}. As $v/[nec_{(\varphi,\tau)}] \subseteq u$, we conclude that $\chi \in u$. Since ψ, χ were generic formulas s.t. $\mathsf{A}(\chi \to \psi) \wedge [im_{(\varphi,\tau)}]\chi \in w$, $w /\!\!/ [im_{(\varphi,\tau)}] \subseteq u$ by the def. of $w /\!\!/ [im_{(\varphi,\tau)}]$, and so $u \in f^c_{top}(\varphi, \tau, w)$ by the def. of f^c_{top}.

5. M^c *satisfies the condition of ability to select answers.*

   This follows immediately from the def. of f^c_{in}.

6. M^c *satisfies the condition of success.*

   Suppose that $v \in f^c_{in}(\varphi, w)$. By the def. of f^c_{in}, there is $w' \in W$ s.t. $wR^c_\Box w'$ and $v \in f^c_{top}(\varphi, \tau, w')$. By the def. of f^c_{top}, $w' /\!\!/ [im_{(\varphi,\tau)}] \subseteq v$. We show that $\varphi \in w' /\!\!/ [im_{(\varphi,\tau)}]$. By $\mathsf{S5_A}$ and Ax2, $\mathsf{A}(\varphi \to \varphi) \wedge [im_{(\tau,\varphi)}]\overline{\varphi} \in w'$. Hence, $\mathsf{A}(\varphi \to \varphi) \wedge [im_{(\tau,\varphi)}]\varphi \in v$ by Ax5, and so $\varphi \in w' /\!\!/ [im_{(\varphi,\tau)}]$. It follows that $\varphi \in v$.

□

The previous lemmas suffice to conclude that VI is strongly complete with respect to the class of all appropriate VI models. In fact, let $\varphi \in \mathcal{L}_{\mathsf{VI}}$ and $\Gamma \subseteq \mathcal{L}_{\mathsf{VI}}$ be such that $\Gamma \nvdash_{\mathsf{VI}} \varphi$. Then, $\Gamma \cup \{\neg\varphi\}$ is consistent, and so it can be extended to a mcs w_0 by Lindenbaum's Lemma. Let M^c be the canonical VI model for w_0. By the definition of W^c and the logic of A, $w_0/\mathsf{A} \subseteq w_0$, and so $w_0 \in W^c$. In addition, by the properties of mcs, $\varphi \notin w_0$. Therefore, for all $\psi \in \Gamma$, $M^c, w_0 \models \psi$ and $M^c, w_0 \not\models \varphi$ by Truth Lemma [Lemma 5.2.20]. Since M^c is an appropriate VI model by Lemma 5.2.21, it follows that φ is not a logical consequence of Γ in the class of appropriate VI models.

5.2.5 Relation with the logic $\mathbf{I}^*$

The language $\mathcal{L}_{\mathbf{I}^*}$ of the logic of imagination $\mathbf{I}^*$ is the fragment of $\mathcal{L}_{\mathsf{VI}}$ without the modalities $\Box$, $[im_{(\varphi,\tau)}]$, and $[nec_{(\varphi,\tau)}]$. The semantics for $\mathcal{L}_{\mathbf{I}^*}$ is based on the notion of *appropriate* $\mathbf{I}^*$ *model*, which is any tuple $\langle W, \mathcal{T}, f_{in}, \nu\rangle$, where $W \neq \varnothing$ is a set of possible states, $\mathcal{T}$ a topic model, $\nu : Prop \to 2^W$ a valuation function, and $f_{in} : \mathcal{L}_{\mathbf{I}^*} \times W \to 2^W$ a selection function satisfying the success condition stated in Definition 5.2.6. The evaluation rules for $\mathsf{A}\varphi$ and $[im_\varphi]\psi$ are as in Definition 5.2.5. The axiom system $\mathbf{I}^*$ is defined by the first four items at the top of Table 5.1, the principles in Proposition 5.2.7 except for $\mathsf{Thm5}$, and the following principle:

$\mathsf{Ax4'}$ $[im_\varphi]\psi \to \mathsf{A}[im_\varphi]\psi$

As shown by Giordani [2019, Theorems 3.4 and 3.8], $\mathbf{I}^*$ is sound and strongly complete with respect to the class of all appropriate $\mathbf{I}^*$ models.

For any $\Gamma \subseteq \mathcal{L}_{\mathbf{I}^*}$ and $\varphi \in \mathcal{L}_{\mathbf{I}^*}$, we write $\Gamma \Vdash_{\mathbf{I}^*} \varphi$ when φ is a logical consequence of Γ in the class of appropriate $\mathbf{I}^*$ models and $\Gamma \Vdash_{\mathsf{VI}} \varphi$ when φ is a logical consequence of Γ in the class of appropriate VI models. In addition, we write $\vdash_{\mathbf{I}^*} \varphi$ when φ is a theorem of $\mathbf{I}^*$ and $\Gamma \vdash_{\mathbf{I}^*} \varphi$ when φ is deducible in $\mathbf{I}^*$ from Γ.

5.2.22. Theorem. *The logic* VI *is a conservative extension of the logic* $\mathbf{I}^*$*: for all* $\Gamma \subseteq \mathcal{L}_{\mathbf{I}^*}$ *and* $\varphi \in \mathcal{L}_{\mathbf{I}^*}$, $\Gamma \Vdash_{\mathbf{I}^*} \varphi$ *iff* $\Gamma \Vdash_{\mathsf{VI}} \varphi$.

Proof:
First, suppose that $\Gamma \Vdash_{\mathbf{I}^*} \varphi$. Since $\mathbf{I}^*$ is complete w.r.t. the class of appropriate $\mathbf{I}^*$ models, $\Gamma \vdash_{\mathbf{I}^*} \varphi$. With the exception of axiom $\mathsf{Ax4'}$, we know already that all axioms of $\mathbf{I}^*$ are theorems of VI and that all inference rules of $\mathbf{I}^*$ are derivable in VI. The proof that $\mathsf{Ax4'}$ is a theorem of VI is as follows:

(1)	$\vdash_{\mathsf{VI}} [im_{\varphi}]\overline{\psi} \rightarrow [im_{(\varphi,\tau)}]\overline{\psi}$	by Thm5
(2)	$\vdash_{\mathsf{VI}} [im_{(\varphi,\tau)}]\overline{\psi} \rightarrow \mathsf{A}[im_{(\varphi,\tau)}]\overline{\psi}$	by Ax4
(3)	$\vdash_{\mathsf{VI}} \mathsf{A}[im_{(\varphi,\tau)}]\overline{\psi} \rightarrow \mathsf{A}\Box[im_{(\varphi,\tau)}]\overline{\psi}$	by Inc1 and logic of A
(4)	$\vdash_{\mathsf{VI}} \mathsf{A}\Box[im_{(\varphi,\tau)}]\overline{\psi} \rightarrow \mathsf{A}[im_{\varphi}]\overline{\psi}$	by $\mathsf{Def}_{[im_{\varphi}]}$ and logic of A
(5)	$\vdash_{\mathsf{VI}} [im_{\varphi}]\psi \rightarrow \mathsf{A}[im_{\varphi}]\psi$	from (1)-(4)

Hence, $\Gamma \vdash_{\mathbf{I}^*} \varphi$ implies $\Gamma \vdash_{\mathsf{VI}} \varphi$. Since VI is sound w.r.t. the class of appropriate VI models [Theorem 5.2.9], we conclude that $\Gamma \Vdash_{\mathsf{VI}} \varphi$. For the other direction, suppose that $\Gamma \nVdash_{\mathbf{I}^*} \varphi$. Then there is an appropriate $\mathbf{I}^*$ model $M = \langle W, \mathcal{T}, f_{in}, \nu\rangle$ and a state $w \in W$ s.t., for all $\psi \in \Gamma$, $M, w \models \psi$ and $M, w \not\models \varphi$. Define $M^* = \langle W^*, \mathcal{T}^*, R^*_{\Box}, f^*_{in}, f^*_{top}, \sim^*, \nu^*\rangle$ by setting:

- $W^* = W$, $\mathcal{T}^* = \mathcal{T}$, $f^*_{in} = f_{in}$, $\nu^* = \nu$;
- $R^*_{\Box} \subseteq W^* \times W^*$ is s.t., for all $w, w' \in W^*$, $wR^*_{\Box}w'$ iff $w = w'$;
- $f^*_{top} : \mathcal{L}_{\mathsf{VI}} \times T \times W^* \rightarrow 2^{W^*}$ is s.t., for all $(\varphi, \tau, w) \in \mathcal{L}_{\mathsf{VI}} \times T \times W^*$, $f^*_{top}(\varphi, \tau, w) = f_{in}(\varphi, w)$;
- $\sim^*$: $\mathcal{L}_{\mathsf{VI}} \times T \rightarrow 2^{W^* \times W^*}$ is s.t., for all $(\varphi, \tau) \in \mathcal{L}_{\mathsf{VI}} \times W^*$ and $w, w' \in W^*$, $w \sim^*_{(\varphi,\tau)} w'$ iff $w = w'$.

It is immediate to check that M^* is an appropriate VI model and that, for all $\psi \in \mathcal{L}_{\mathbf{I}^*}$ and w in W, $M, w \models \psi$ iff $M^*, w \models \psi$. Therefore, for all $\psi \in \Gamma$, $M^*, w \models \psi$ and $M^*, w \not\models \varphi$, and so $\Gamma \nVdash_{\mathsf{VI}} \varphi$, as desired. □

The proof of Theorem 5.2.22 shows that the logic of imagination $\mathbf{I}^*$ can be obtained from the logic of voluntary imagination VI by avoiding modeling the control of the agent over her imagination acts(as represented by $R_{\Box}$), the process of topic selection (as represented by f_{top}), and the connection between input-topic pairs and partitions of W (as represented by $\sim$).

5.3 Back to the key questions

In this section, we consider some relevant issues connected to our three initial questions concerning the logic, epistemic value, and voluntary components of ROMS. A more extensive and systematic study of the implications of our framework is left for future work.

5.3.1 What is the logic of imagination as ROMS?

In Sections 5.2.3 and 5.2.4 we have seen that the imagination operators $[im_\varphi]$ and $[im_{(\varphi,\tau)}]$ have a number of promising properties. They are non-normal, in accordance with the fact that, when we imagine something, we do not imagine all logical consequences of what we are imagining (when we imagine that Holmes is talking to Watson, we do not imagine that Holmes is talking to Watson and either there is or there is not a seminar at the ILLC today). In addition, they are non-monotonic, in accordance with the fact that, given different inputs, we import different information into the imagined scenario (if the input is *The paper has been accepted to DEON*, we imagine ourselves traveling to the conference, but, if the input is *The paper has been accepted to DEON and there is a global pandemic*, we do not). Finally, they satisfy a number of closure principles, i.e., $\mathsf{K}_{[im_\varphi]}$, $\mathsf{K}_{[im_{(\varphi,\tau)}]}$, $\mathsf{C}_{[im_\varphi]}$, $\mathsf{C}_{[im_{(\varphi,\tau)}]}$, $\mathsf{Ax5}$, and $\mathsf{Thm4}$. Of these, the principles of Conjunction Introduction $\mathsf{C}_{[im_\varphi]}$ and $\mathsf{C}_{[im_{(\varphi,\tau)}]}$ may be controversial. As Berto [2018, pp. 1879-80], building on a famous example from Quine [1960, p. 222], points out,

> The explicit input indexing $[im_\varphi]$ involves Caesar being in command of the US troops in the war of Korea. We can imagine him using bombs, ψ_1, importing in the representation the weapons available in the Korean war, or we can imagine him using catapults, ψ_2, importing the military apparatus available to Caesar. However, one would not thereby infer $[im_\varphi](\psi_1 \wedge \psi_2)$, Caesar's employing both bombs and catapults. One can imagine *that*, too, if one likes, but it should not come out as an automatic entailment from the logic of imagining. [Notation adapted.]

Berto's [2018] solution to this issue is based on the fact that acts of imagination are contextually determined. This means that the same explicit input can trigger different acts of imagination in different contexts. In the example adapted from Quine, the conclusion that, given input φ, we imagine Caesar's employing both bombs and catapults is reached due to an obvious contextual shift. But once the context is fixed, principles of Conjunction Introduction are intuitive. Formally, a way to implement this solution is by indexing imagination operators with variables for contexts. This is a mere suggestion in Berto [2018]. Here's how we can make it work in our framework: Caesar employing bombs and Caesar employing catapults are different cells in the partition connected with input *Caesar is in command of the US troops in the war of Korea* and topic *Which weapons does Caesar use?*. Let this topic be τ. The context in which we imagine Caesar's using bombs is a situation in which the following is true:

$$[im_{(\varphi,\tau)}]\psi_1 \wedge \Diamond[im_{(\varphi,\tau)}]\psi_2$$

That is, we select cells in the τ-partition for φ where Caesar uses bombs (ψ_1), even if, in a different context, we could have selected cells where he uses catapults (ψ_2). Since

$$[im_{(\varphi,\tau)}]\psi_1 \wedge \Diamond[im_{(\varphi,\tau)}]\psi_2 \rightarrow [im_{(\varphi,\tau)}](\psi_1 \wedge \psi_2)$$

is *not* a valid principle, the inference to $[im_{(\varphi,\tau)}](\psi_1 \wedge \psi_2)$ is blocked. This suggests that reference to topics is crucial to provide a correct analysis of the logic underlying imagination acts.

5.3.2 How do ROMS relate to knowledge?

There are at least two properties that imagination should have to be a vehicle to gain new knowledge: first, it should be possible for imagination to be *selective*; second, it should be possible to *learn from imagination*. As to the first property, we have seen in Section 5.1 that acts of ROMS are constrained both by what we know or believe and by the goal for which they are pursued. Consider the following situation, discussed by Williamson [2016, p. 114], concerning a group of our ancestors who suppose that there are wolves in the forest they are about to enter:

> To serve that purpose well, the imagination must be both selective and reality-oriented. They [the ancestors] could imagine the wolves bringing them food to eat, but doing so would be a waste of time, and a distraction from more practically relevant possibilities. An imagination that clutters up the mind with a bewildering plethora of wildly unlikely scenarios is almost as bad as no imagination at all. It is better to have an imagination that concentrates on fewer and more likely scenarios. One's imagination should not be completely independent of one's knowledge of what the world is like.

In our framework, we can represent how imagination is selective and reality-oriented. On the one hand, oriented selectivity has to do with the fact that the selected scenarios are to be close to what we take the actual world to be like: function f_{in} is introduced to do this job. On the other hand, oriented selectivity has also to do with the fact that the selected scenarios are to be consistent with a specific issue, or question, addressed in the exercise: this is captured by the selection of the topic operated by f_{top}.

As to the second property – the possibility of learning from imagination – the problem is how the imagined scenario can be specified in an epistemically legitimate way. As we saw in Section 5.1, Williamson [2016, p. 116] takes the key to be the combination of voluntary and involuntary components in acts of imagination:

> [H]owever difficult the jump, one can imagine succeeding with it, and however easy the jump, one can imagine failing with it. How can one

> learn anything relevant from what one chooses to imagine? Such incomprehension indicates neglect of the distinction [...] between voluntary and involuntary exercises of the imagination. When the hunter makes himself imagine trying to jump the stream, his imagination operates in voluntary mode. But he neither makes himself imagine succeeding nor makes himself imagine failing. [...] He imagines the antecedent of the conditional voluntarily, the consequent involuntarily. Left to itself, the imagination develops the scenario in a reality-oriented way, by default.

In our framework, this view can be represented through propositions like:

Learning 1: $\neg[im_{\varphi}]\psi \wedge [im_{(\varphi,\tau)}]\psi$

Learning 2: $\neg[im_{\varphi}]\psi \wedge [im_{(\varphi,\tau)}]\psi \wedge [im_{\varphi}]([nec_{(\varphi,\tau)}]\psi \vee [nec_{(\varphi,\tau)}]\neg\psi)$

In *Learning 1* and *Learning 2*, $\neg[im_{\varphi}]\psi$ states that what the agent imagines given input φ is consistent with $\neg\psi$; that is, the truth value of ψ can vary based on the choice and specification of a topic. The agent is thus free to opt for a topic and a specific point of view on it, corresponding to one or more cells in the partition it determines. Still, *once* that point of view is selected, the consequences of the selection – the propositions that come out to be true given the selection – are no longer in control of the agent: the antecedent and the cell in the partition can be voluntarily chosen, but the consequent – what is implied by that cell – is involuntarily settled. This is captured by the second conjunct in *Learning 1* and *Learning 2*: $[im_{(\varphi,\tau)}]\psi$ states that, given the selected cells of the τ-partition for φ, $\neg\psi$ is no longer consistent with what the agent imagines, that the specified imagined scenario forces the truth of ψ. For the last conjunct in *Learning 2*, it is useful to have a look at the semantics first. For any state w in an appropriate VI model M, let $[w]_{(\varphi,\tau)} = \{w' \in W \mid w \sim_{(\varphi,\tau)} w'\}$ be the equivalence class of w under $\sim_{(\varphi,\tau)}$. By Definition 5.2.5, we have:

$M, w \models [im_{\varphi}]([nec_{(\varphi,\tau)}]\psi \vee [nec_{(\varphi,\tau)}]\neg\psi)$ iff

(i) for all $v \in f_{in}(\varphi, w)$, either $M, v \models [nec_{(\varphi,\tau)}]\psi$ or $M, v \models [nec_{(\varphi,\tau)}]\neg\psi$
iff for all $v \in f_{in}(\varphi, w)$, either $[v]_{(\varphi,\tau)} \subseteq [\![\psi]\!]$ or $[v]_{(\varphi,\tau)} \subseteq [\![\neg\psi]\!]$

(ii) $t([nec_{(\varphi,\tau)}]\psi \vee [nec_{(\varphi,\tau)}]\neg\psi) \sqsubseteq t(\varphi)$
iff $t(\varphi) \oplus t(\psi) \sqsubseteq t(\varphi)$
iff $t(\psi) \sqsubseteq t(\varphi)$

Accordingly, $[im_{\varphi}]([nec_{(\varphi,\tau)}]\psi \vee [nec_{(\varphi,\tau)}]\neg\psi)$ states that (i) once input φ is received, every complete specification of the basic imagined scenario relative to

topic τ determines the truth value of ψ and that (ii) ψ is on-topic relative to φ. More technically, according to (i), the proposition $[\![\psi]\!]$ restricted to the basic imagined scenario $f_{in}(\varphi, w)$ is a union of cells in the τ-partition for φ (and the same is true for $[\![\neg\psi]\!]$). Following Lewis [1988] [see also footnote 11], this can be understood as saying that ψ is about the questions raised by topic τ given input φ, or that ψ is relevant to issue τ. Hence, while *Learning 1* captures the idea that ψ can be learned, or discovered, in virtue of the selected topic specification, *Learning 2* also keeps track of the fact that the learned proposition is relevant to that very topic.[15]

5.3.3 What is voluntary in a ROMS?

We have provided an analysis of the voluntary and involuntary components of imagination acts in Section 5.1 and indicated how our logic VI represents voluntary components in various parts of Section 5.2. The following list summarizes how $\mathcal{L}_{\mathsf{VI}}$ can be used to describe a number of features of imagination acts related to these components.

1. Selection of input and topic:

   1.1. $\langle im_{\varphi}\rangle\,\overline{\varphi}$ says that the agent actually entertains an imagination act based on input φ;

   1.2. $\langle im_{(\varphi,\tau)}\rangle\,\overline{\varphi}$ says that the agent actually processes input φ in light of topic τ.

2. Possibility to select non-selected inputs and topics:

   2.1 $\neg\langle im_{\varphi}\rangle\,\overline{\varphi} \wedge \mathsf{E}\langle im_{\varphi}\rangle\,\overline{\varphi}$ says that the agent does not actually entertain an imagination act based on input φ even if, in different circumstances (e.g., when input φ is received), she would entertain an act based on such input;

   2.2 $\neg\langle im_{(\varphi,\tau)}\rangle\,\overline{\varphi} \wedge \Diamond\langle im_{(\varphi,\tau)}\rangle\,\overline{\varphi}$ says that the agent does not actually process input φ in light of topic τ even if she has the option to do so.

3. Deliberativeness in the selection and specification of a topic:

   3.1 $\langle im_{(\varphi,\tau)}\rangle\,\overline{\varphi} \wedge \neg\Box\langle im_{(\varphi,\tau)}\rangle\,\overline{\varphi}$ says that the agent actually processes input φ in light of topic τ, but it is not settled that she does so;

[15] Strictly speaking, both *Learning 1* and *Learning 2* only represent what we *could learn* when we entertain an imaginative act. In fact, in order to represent what we *actually learn*, we should include, besides a proposition stating that ψ holds at all the scenarios whose selection is based on input φ with topic τ, a proposition stating that ψ holds in the φ-worlds that are closest to the actual world. In other terms, we should supplement the epistemic conditional with a corresponding ontic conditional, thus grounding the truthfulness of our epistemic act.

3.2 $\langle im_{(\varphi,\tau)} \rangle \overline{\varphi} \wedge [im_{(\varphi,\tau)}]\psi \wedge \Diamond(\langle im_{(\varphi,\tau)} \rangle \overline{\varphi} \wedge \langle im_{(\varphi,\tau)} \rangle \neg\psi)$ says that, given input φ and the selected specification of topic τ, the agent actually imagines that ψ, even if it is not settled that the agent specifies the topic in a way that makes her imagine that ψ.

Finally, if we assume that the set of topics is finite, we can use the formula

$$\langle im_{\varphi} \rangle \overline{\varphi} \wedge \bigwedge_{\tau \in T} [im_{(\varphi,\tau)}] \neg \overline{\varphi}$$

to express that the agent actually entertains an imagination act based on input φ, but chooses not to process it in light of any topic.

5.4 Conclusion

In this chapter, we have claimed that imagination as ROMS has a number of features: it is agentive and episodic; it starts from a deliberate input, an initial supposition concerning what is to be mentally simulated; it integrates such input on the basis of the agent's background knowledge and beliefs, without mobilizing all of them – rather, only those deemed relevant to the topic of the imaginative act; and it has a purpose or goal, which is to address some question, the answer to which drives the agent's interests. We have argued that some of these components of an exercise of ROMS (e.g., processing an initial input, choosing the goal) involve voluntary choices on the side of the agent, while others are involuntary and automatic (e.g., retrieving information from one's background knowledge and beliefs to integrate the input). We have then presented a sound and complete logic of voluntary imagination (VI), characterized by modal operators expressing the imaginative options open to an agent, what is imagined given an input, what is imagined and what is necessary given an input and a topic. The semantics combines ideas from STIT semantics and from semantics for counterfactuals with a mereology of topics from aboutness or subject matter semantics.

VI allows us to express and address issues concerning (what we claim to be) three main, interconnected questions concerning imagination as ROMS: (I) *What is its logic?* We have shown that the imagination operators of VI have some noteworthy closure properties, in spite of their being non-normal and non-monotonic modals (in particular, given an input, or an input and a topic, one doesn't imagine all the logical consequences of the input itself). (II) *How does imagination as ROMS relate to knowledge?* We have shown how one can express in VI the conditions under which an agent can learn something new via an act of ROMS. This relates rather strictly to question (III), *What is voluntary and what is not in ROMS?* VI can express the distinction between voluntary and involuntary components, and thus help to make formally precise the idea, entertained by various authors, that imagination can allow us to gain new knowledge because some aspects of mental simulation are not arbitrary, but governed by the automatic and

generally reliable mechanisms that regulate the administration and revision of our beliefs in the light of new information.

One main direction of further investigation within the proposed framework, flagged since Section 5.1, involves the temporal dynamics of ROMS: how episodes of mental simulation develop temporally, while representing actions and events that themselves unfold in time. We have left this issue for future work, while conjecturing that combining the ideas developed in Chapter 4 with the techniques of Dynamic Epistemic Logic may help with it.

Part Two

Norms

Chapter 6

From ideal to actual prescriptions in dynamic deontic logic

Ascribing responsibility is not only a matter of identifying who caused a certain result (*actus reus* question) and why they did it (*mens rea* question), as we saw in Chapter 3, but also of determining which actions *ought* and *ought not* to be done: when we hold someone responsible for something, it is partly because that person transgressed some moral or legal norm, because they did something "wrong." In this part of the dissertation, we study deontic logics to reason about the senses in which doing something can be "wrong." Violating a norm is one of such senses. But what if a norm has already been violated? And what if it is not possible to avoid violating some norm because of the circumstances or of an intrinsic defect in the normative system? What ought to be done in such cases?

By merging insights from different traditions in deontic logic, in this chapter we design a dynamic deontic system in the tradition of Meyer [1988] in which we can distinguish different levels of "wrongfulness" and analyze the interaction between them. Our main contribution is the formulation of a rich deontic classification of states, actions, and sequences of action, which allows us to introduce deontic operators expressing what we will call *actual prescriptions*, i.e., prescriptions that are sensitive to what can actually be done in a given situation. We use simple real-life examples to show that the new operators have desirable properties, interact effectively with standard deontic operators expressing other important kinds of prescriptions, and are not affected by paradoxes of dynamic deontic logic concerning sequences of actions [Anglberger, 2008; van der Meyden, 1996].

Outline. We start, in Section 6.1, with a concise overview of the standard deontic systems and paradoxes that inspired our proposal. The reader who is familiar with Standard Deontic Logic (SDL), Chisholm's [1963] paradox, and Meyer's [1988] Propositional Dynamic Deontic Logic ($\mathrm{PD}_e\mathrm{L}$) should feel free to jump to the end of Section 6.1.2, where we discuss Angleberger's [2008] paradox and van der Meyden's [2008] paradox to partly motivate our proposal. In Section

6.2, we explain and support the concepts and assumptions on which our proposal is based. We present our dynamic deontic logic with optimality ($\mathsf{PD}_e\mathsf{LO}$) in Section 6.3: the syntax and semantics are introduced in Section 6.3.1 and a sound and complete axiomatization is provided in Section 6.3.2. In Section 6.4, we use the resources of $\mathsf{PD}_e\mathsf{LO}$ to define four deontic categories of states and actions and to introduce deontic operators expressing prescriptions at four different normative levels. These include new operators for actual prescriptions that apply to one-step actions [Section 6.4.1] and to sequences of actions [Section 6.4.2]. We discuss how the new operators can be used to analyze simple norm-governed transition systems as we introduce them. Section 6.5 summarizes the main results.

This chapter is based on Canavotto and Giordani [2019], which develops ideas first presented in Giordani and Canavotto [2016].

6.1 Background and motivations

The main aim of this section is to provide the background to situate our work in deontic logic. We start by presenting the standard deontic logic SDL and briefly discuss two developments stemming from Chisholm's [1963] paradox. We then introduce Meyer's [1988] "dynamic" solution to the paradox and explain how new paradoxes arising in his system led us to move to a richer framework.

6.1.1 SDL, ideality, and contrary-to-duties

SDL is a modal logic with modal operators O for obligation and P for permission. Formulas like $O\varphi$ mean "φ is obligatory" or "it ought to be that φ" and formulas like $P\varphi$ mean "φ is permitted" or "it may be that φ." The two operators work, respectively, as the box and the diamond of a normal modal logic, and are thus interdefinable: if O is assumed as the primitive operator, $P\varphi$ abbreviates $\neg O\neg\varphi$; similarly, if P is assumed as the primitive operator, $O\varphi$ abbreviates $\neg P\neg\varphi$. Accordingly, φ is permitted if its negation is not obligatory and it is obligatory if its negation is not permitted. Formulas like $F\varphi$, which mean "φ is forbidden," are introduced as abbreviations of $\neg P\varphi$ (or, equivalently, $O\neg\varphi$). So, φ is forbidden if it is not permitted (or, equivalently, if its negation is obligatory).

SDL has a standard possible world semantics. Models are tuples

$$M = \langle W, R_O, \nu \rangle$$

where $W \neq \varnothing$ is a set of possible states, $\nu : Prop \to 2^W$ is a valuation function, and $R_O \subseteq W \times W$ is a deontic accessibility relation, which is assumed to be serial (for every state w, there is a state w' such that wR_Ow'). Intuitively, the accessibility relation R_O relates every state w to its *ideal alternatives*, that is, to those states where every obligation holding at w is complied with. Under this

reading, the seriality of R_O amounts to the requirement that every possible state is governed by a consistent set of obligations. The semantics of the operators O and P is defined by the standard evaluation rules:

$$M, w \models O\varphi \quad \text{iff} \quad \text{for all } w' \in W, \text{ if } wR_Ow', \text{ then } M, w' \models \varphi$$
$$M, w \models P\varphi \quad \text{iff} \quad \text{there is } w' \in W \text{ such that } wR_Ow' \text{ and } M, w' \models \varphi$$

That is, $O\varphi$ is true at w if φ is true at all ideal alternatives to w and $P\varphi$ is true at w if φ is true at some ideal alternative to w. It follows from standard results in modal logic [Blackburn et al., 2001, Chapter 4.2] that SDL is axiomatized by the axioms and rules of the normal modal logic KD for O.

There is a general consensus in the literature that SDL suffers from a number of paradoxes that hardly make it an appropriate logic for deontic reasoning.[1] One of the most serious difficulties has to do with contrary-to-duty obligations (henceforth: CTDs), which concern what ought to be done in case another, primary obligation has been violated – if we ought to be punctual (primary obligation), then the obligation that we ought to apologize if we are not punctual is a CTD obligation. There are a number of paradoxes centering around CTDs [see Carmo and Jones, 2002 and Hilpinen and McNamara, 2013, Section 8.5]. Chisholm's [1963] paradox is the one that best shows that SDL does not have the resources to handle them. Here is the problem. Consider the following sentences:

(a) We ought not to be robbed.

(b) If we are not robbed, we ought not to call the police.

(c) If we are robbed, we ought to call the police.

(d) We are robbed.

Deontic logicians agree that, intuitively, the Chisholm's set consisting of the four sentences (a) to (d) is consistent and that its members are logically independent. But no formalization of the four sentences in SDL meets both requirements.[2]

[1] This does not mean that SDL is not an important system. As Hilpinen and McNamara [2013, p. 39] have it: "[SDL] is hardly a widely popular system of logic with only occasional outliers rejecting it as the title might suggest. Rather, it is the most widely known, well-studied system, and central in the accelerated historical development of the subject over the last 50 or so years. As such, it serves as a historical comparator, where various important developments in the subject were explicit reactions to its perceived shortcomings, and even when not, sometimes can be fruitfully framed as such."

[2] There are two ways to formalize a conditional obligation "if p it ought to be that q" in SDL: (1) $O(p \to q)$ and (2) $p \to Oq$. Granted that sentences (a) and (d) in the Chisholm's set are formalized as $O\neg r$ and r respectively, it is not difficult to see that if (b) is formalized as $O(\neg r \to \neg c)$ and (c) is formalized as $r \to Oc$, then $Oc \wedge \neg Oc$ is deducible in SDL from the four sentences. If (a) is formalized using (2) and (b) using (1) or if both sentences are formalized using either (1) or (2), then the members of the set turn out not to be logically independent. See Hilpinen and McNamara [2013, p. 85] for more details.

Looking at the semantics, the source of the problem is that in models for SDL what is obligatory is determined by how things are at *ideal* states where the law is completely complied with. But CTDs concern how things are at *non-ideal* states where some laws have been violated. As Lewis [1974, p. 1] puts it,

> It ought not to be that you are robbed. A fortiori, it ought not to be that you are robbed and then helped. But you ought to be helped, given that you have been robbed. The robbing excludes the best possibilities that might otherwise have been actualized, and the helping is needed in order to actualize the best of those that remain. Among the possible worlds marred by the robbing, the best of a bad lot are some of those where the robbing is followed by helping.

Let r stand for the sentence "we are robbed" and h for the sentence "we are helped." The point is that, if $O\neg r$ is true at a state w, then all ideal alternatives to w are states where r (hence $r \wedge h$) is false. But, as Lewis notices, supposing that r is true, what is obligatory should be determined by (non-ideal) states where r rather than $\neg r$ is true and, in particular, by the *best* such states (where, presumably, $r \wedge h$ rather than $r \wedge \neg h$ is true). In other words, *given a breach of the law, ideality seems to be too high a standard.*

This diagnosis has led to two main "semantic approaches" to CTDs:[3]

1. In preference-based semantics [Hansson, 1969; Lewis, 1974] the accessibility relation R_O is replaced with a betterness ordering between states. This makes it possible to select both the best absolute states (e.g., the ideal states where one is neither robbed nor helped) and the best states where some fact ψ occurs (e.g., the non-ideal where one is robbed but helped). Best absolute states are used to determine *unconditional obligations* represented by formulas like $O\varphi$, while the best ψ-states are used to determine *conditional obligations* represented by formulas like $O(\varphi \,|\, \psi)$ (read: "given ψ, it ought to be that φ").[4] If r stands for "we are robbed" (as before) and c stands for "we call the police," the Chisholm's set is then given the following representation, which can be shown to satisfy the requirements of consistency and logical independence:

   (a) $O\neg r$, (b) $O(\neg c \,|\, \neg r)$, (c) $O(c \,|\, r)$, (d) r.

[3] What follows is a concise presentation of key ideas that are useful to contextualize our work, not an introduction. More about the first approach can be found in Prakken and Sergot [1996, 1997] and more about the second can be found in Carmo and Jones [2002]. For completeness, we should also mention that there is a family of "syntactic approaches" to CTDs and, more generally, to conditional obligations, which include, e.g., input-output logics [Makinson and van der Torre, 2000, 2001; Parent and van der Torre, 2013], default logics [Horty, 2012], and logics of sequential obligations [Governatori and Rotolo, 2006; Governatori et al., 2016].

[4] As usual in conditional logics, $O\varphi$ is introduced as an abbreviation for $O(\varphi \,|\, \top)$.

2. In semantics based on so-called sub-ideality [Jones and Pörn, 1985; Carmo and Jones, 1995], SDL models are supplemented with a distinction between absolutely ideal states and ideal versions of individual states, where an ideal version of a state may be a "sub-ideal" state where some obligation is violated. In Lewis' example, the absolutely ideal states are those where $r \wedge h$ is false and the ideal versions of an r-state are the sub-ideal states where $r \wedge h$ is true. While absolutely ideal states are used to determine *ideal obligations* represented by formulas like $O_{\mathrm{i}}\varphi$ (read: "*ideally*, it ought to be that φ"), ideal versions of a state are used to determine *actual obligations* represented by formulas like $O_{\mathrm{a}}\varphi$ (read: "*given the circumstances*, it ought to be that φ"). Given an appropriate necessity operator $\blacksquare$, the Chisholm's set can then be satisfactorily[5] formalized as follows:

   (a) $O_{\mathrm{i}}\varphi$, (b) $\blacksquare(\neg r \rightarrow O_{\mathrm{a}}\neg c)$, (c) $\blacksquare(r \rightarrow O_{\mathrm{a}}c)$, (d) r.

The previous approaches to CTDs center around the idea that, in case the law is violated, one ought to realize "the best [alternatives] of a bad lot." But there is another suggestion in the Lewisian analysis above, namely that CTDs are obligations that are triggered by the performance of a prohibited action.[6] Going back to Lewis' example, although it is true that, *ideally*, it ought to be that neither the action *robbing* nor the sequence of actions *robbing-and-then-helping* are performed, *after* the action *robbing* is performed, performing the action *helping* becomes obligatory. In other words, the primary obligation not to rob and the CTD obligation to help characterize different states: the initial state and the end state of an individual action of type *robbing*. This line of reasoning is at the heart of the "dynamic approach" to CTDs proposed in the context of Propositional Dynamic d*e*ontic Logic (PD$_e$L) [Meyer, 1988].

6.1.2 PD$_e$L, process norms, and where we are headed

PD$_e$L, introduced by Meyer [1988], is obtained by extending a version of PDL [Harel et al., 2000; see also Chapter 2.3.2] with deontic elements. As the language of PDL, the language of PD$_e$L is characterized by two categories of expressions: (names of) action types and formulas. Action types are built from a set *Atm* of atomic types using the operators ; of sequential composition, $\cup$ of non-deterministic composition, $\cap$ of parallel composition, and $\bar{\cdot}$ of action negation.[7]

[5] "Satisfactorily" means that the set meets the requirements of consistency and logical independence. See Carmo and Jones [1997, 2002] for a discussion of the limits of Jones and Pörn [1985] and Carmo and Jones [1995] and a more elaborate framework.

[6] Most CTDs do indeed have this property, but see Prakken and Sergot [1996, 1997] for examples of CTDs that, on the face of it, do not involve actions or the passage of time. We will ignore these examples in what follows.

[7] The language of PD$_e$L also includes an action constant $\varnothing$ denoting the impossible action and an operator to build so-called conditional actions. We omit them from the presentation because they do not play any role in what follows.

Intuitively, where α and β are action types, we have that:

1. $\alpha;\beta$ is the type instantiated by any token instantiating α and β in sequence;
2. $\alpha \cup \beta$ is the type instantiated by any token instantiating either α or β;
3. $\alpha \cap \beta$ is the type instantiated by any token instantiating α and β in parallel;
4. $\overline{\alpha}$ is the type instantiated by any token that does not instantiate α.

Besides dynamic formulas like $[\alpha]\varphi$ and $\langle\alpha\rangle\,\varphi$ (meaning, respectively, "doing α necessarily results in a φ-state" and "doing α possibly results in a φ-state"), the featured formula of PD$_e$L is a propositional constant *id*, which means "all obligations are complied with" or "the present state is ideal."[8] Deontic operators for obligation, permission, and prohibition applying to action types can then be introduced by means of the following abbreviations, inspired by Anderson [1958] and Kanger [1957]:

$$P'\alpha := \langle\alpha\rangle\, id \qquad F'\alpha := \neg P'\alpha \qquad O'\alpha := \neg P'\overline{\alpha}$$

Hence, it is permitted to do α if doing α possibly results in an ideal state; it is forbidden to do α if it is not permitted to do it; finally, it is obligatory to do α if it is not permitted to refrain from doing it (equivalently, if refraining from doing it is forbidden).

The semantics for PD$_e$L enriches the semantics for PDL. Models for PD$_e$L are tuples $M = \langle W, R, \mathbf{ideal}, \nu\rangle$, where $W \neq \varnothing$ is a set of possible states, $\nu : Prop \to 2^W$ is a valuation function, $R : Types \to 2^{W\times W}$ assigns to every action type α an accessibility relation R_α that relates every state w with the possible outcomes of α at w (if any), and $\mathbf{ideal} \subseteq W$ is a set of ideal states where no law is violated. The evaluation rules for the propositional constant *id* and the dynamic modalities $[\alpha]$ are as expected:

$$\begin{array}{lll} M, w \models id & \text{iff} & w \in \mathbf{ideal} \\ M, w \models [\alpha]\varphi & \text{iff} & \text{for all } w' \in W, \text{ if } wR_\alpha w', \text{ then } M, w' \models \varphi \end{array}$$

The semantics for the deontic operators is then as follows:[9]

$$\begin{array}{lll} M, w \models P'\alpha & \text{iff} & R_\alpha(w) \cap \mathbf{ideal} \neq \varnothing \\ M, w \models F'\alpha & \text{iff} & R_\alpha(w) \cap \mathbf{ideal} = \varnothing \\ M, w \models O'\alpha & \text{iff} & R_{\overline{\alpha}}(w) \cap \mathbf{ideal} = \varnothing \end{array}$$

[8] Actually, Meyer [1988] introduces a propositional constant *vio* that means "some obligation is violated." Since *vio* can be defined as $\neg id$ and *id* as $\neg vio$, this difference is immaterial. We use *id* for uniformity with our proposal.

[9] As usual, for any binary relation R on a set W and $w \in W$, $R(w) = \{w' \in W \mid wRw'\}$.

(CPL)	All tautologies of CPL	(MP)	From $\varphi \to \psi$ and φ, infer ψ
($\mathsf{K}_{[\alpha]}$)	The axiom schema of K for $[\alpha]$	($\mathsf{RN}_{[\alpha]}$)	From φ, infer $[\alpha]\varphi$
	(I) Axioms for $;\ \cup\ \cap$		**(II) Axioms for** $\overline{\ }$
(Seq)	$[\alpha;\beta]\varphi \leftrightarrow [\alpha][\beta]\varphi$	(Neg1)	$[\overline{\alpha;\beta}]\varphi \leftrightarrow [\overline{\alpha}]\varphi \wedge [\alpha;\overline{\beta}]\varphi$
(Com)	$[\alpha \cup \beta]\varphi \leftrightarrow [\alpha]\varphi \wedge [\beta]\varphi$	(Neg2)	$[\overline{\alpha}]\varphi \to [\overline{\alpha \cup \beta}]\varphi$
(Par)	$[\alpha]\varphi \to [\alpha \cap \beta]\varphi$	(Neg3)	$[\overline{\alpha \cap \beta}]\varphi \leftrightarrow [\overline{\alpha}]\varphi \wedge [\overline{\beta}]\varphi$
		(Neg4)	$[\overline{\overline{\alpha}}]\varphi \leftrightarrow [\alpha]\varphi$

Table 6.1: The axiom system $\mathrm{PD}_e\mathrm{L}$

Accordingly, doing α is: permitted if some α-transition ends in an ideal state, forbidden if every α-transition ends in a non-ideal state, and obligatory if every $\overline{\alpha}$-transition ends in a non-ideal state.[10] The axiom system $\mathrm{PD}_e\mathrm{L}$ is defined by the axioms and rules in Table 6.1, which are sound with respect to the semantics presented by Meyer [1988].[11]

As convincingly argued by Meyer [1988], $\mathrm{PD}_e\mathrm{L}$ provides an interesting perspective on many paradoxes of SDL. In particular, the idea that CTDs are triggered by the performance of a prohibited action can be naturally expressed by formulas like $O'\overline{\alpha} \wedge [\alpha]O'\beta$. So, letting ρ be the action type *robbing* and κ the action type *calling the police*, the first three sentences of the Chisholm's set can be formalized as follows: (a) $O'\overline{\rho}$, (b) $[\overline{\rho}]O'\overline{\kappa}$, (c) $[\rho]O'\kappa$.[12] It can be easily proved that, in $\mathrm{PD}_e\mathrm{L}$, these imply

$$O'\overline{\rho} \wedge O'(\overline{\rho};\overline{\kappa}) \wedge \neg O'(\rho;\kappa),$$

which nicely shows that the system distinguishes CTDs like the one in (c) from ideal, compliant-with-duty obligations like the one in (b): while the latter direct to the performance of an ideal course of action, the former do not.

Unfortunately, however, $\mathrm{PD}_e\mathrm{L}$ is subject to two serious paradoxes:

Angleberger's paradox. [Anglberger, 2008] Given the principles on action negation assumed by Meyer [1988, p. 113], the formula

$$F'\alpha \to [\alpha]F'\beta \qquad (\mathsf{Ang})$$

[10] Recall that an α-transition is a pair (w, w') such that $wR_\alpha w'$.

[11] To our knowledge, no completeness result for Meyer's original system has been presented in the literature.

[12] Like PDL, $\mathrm{PD}_e\mathrm{L}$ is about *potential* rather than *actual* agency [cf. Section 2.3.2]. As a consequence, it lacks the resources to express the last sentence in the Chisholm's set, i.e., "we are robbed." There are different ways to refine frameworks like $\mathrm{PD}_e\mathrm{L}$ to overcome this problem [see, e.g., Herzig et al., 2018 and Broersen, 2003, Chapter 5].

is a theorem of $\mathrm{PD}_e\mathrm{L}$. Ang says that, after performing a prohibited action, every action becomes prohibited: there is nothing that can be done to make up for a breach of the law. Besides trivializing the concept of a CTD obligation, the derivability of Ang shows that the first three members of the Chisholm's set, as formalized by Meyer, are not logically independent, as (a) turns out to imply (c). An obvious potential solution to this problem is to extend $\mathrm{PD}_e\mathrm{L}$ with the following version of axiom D:

$$P'\alpha \vee P'\overline{\alpha} \qquad (\mathsf{D}_{P'})$$

which expresses that some action is permitted. But if $\mathsf{D}_{P'}$ is added to $\mathrm{PD}_e\mathrm{L}$, then the formula $F\alpha \to [\alpha]\,\bot$, which says that no prohibited action can be performed, becomes derivable [Anglberger, 2008, p. 432].

van der Meyden's paradox. [van der Meyden, 1996] Given the definition of the operator P', the formula

$$\langle\alpha\rangle\, P'(\beta) \to P'(\alpha;\beta) \qquad (\mathsf{vDM})$$

is a theorem of $\mathrm{PD}_e\mathrm{L}$. But vDM has the following implausible instance: "If there is a way to steal money after which it is permitted to make a call, then it is permitted to steal money and then make a call."

Angleberger's paradox has not received much attention in the literature. On the one hand, since the proof of Ang depends essentially on Meyer's characterization of the operator of action negation – which is controversial for independent reasons [Broersen, 2003, 2004] – it is fair to think that the paradox is not deontic in nature. On the other hand, the validity of Ang suggests that a distinction between ideality and sub-ideality has still a role to play in dynamic deontic logic: There are situations in which, after not complying with some obligation, we cannot avoid violating some other obligation, no matter what we do. (For instance, we all know that, on the day of the submission deadline, after being late for the meeting with our co-authors, we will inevitably be late for the seminar.) If ideality is the only standard to determine what is permitted or obligatory, then no reasonable notion of permission or obligation will guide our behavior in such situations. But, intuitively, *there is* a reasonable, non-trivial obligation we should comply with: "to make the best out of the sad circumstances" [Hansson, 1969, p. 395]. As we will see in the next Section 6.2, the distinction between ideality and sub-ideality also turns out to be crucial to distinguish different important categories of non-compliant behavior.

Turning to vDM, the paradox derives from the fact that, according to Meyer's definitions of the deontic modalities, the deontic status of a course of action is completely determined by the deontic status of its possible outcomes, while what happens during its execution is irrelevant. Thus, borrowing an expression from Broersen [2003], Meyer's deontic operators express *goal norms* rather than *process norms*. Van der Meyden's [1996] solution is to model process norms

by building a deontic logic on top of Pratt's [1979] process logic (rather than PDL). This framework allows to describe the properties of different segments of a transition. The idea is then to supplement process logic with a distinction between "green" and "red" transitions, where a transition is green when all of its segments are "[normatively] good" [van der Meyden, 1996, p. 467] or, in our terminology, *ideal*. This distinction is used to provide a semantics for two operators of process-permission and an operator of process-prohibition. Van der Meyden [1996] leaves the problem of introducing an appropriate operator for process-obligation unresolved.

In order to address the above-mentioned issues, we will design a dynamic deontic system incorporating a distinction between ideal and sub-ideal states and a corresponding distinction between ideal and sub-ideal actions. In addition, we will provide a characterization of process- permissions, prohibitions, and obligations that is sensitive to these distinctions. We introduce and further motivate the key ideas underlying our framework in the next section. We anticipate that, taken individually, the semantical ingredients characterizing it are common in the deontic logic literature we briefly surveyed in this section. The novelty lies in the specific way in which they are combined and used to model process norms.

In this regard, some connections with the most recent literature on dynamic deontic logic should be mentioned. First, modalities for process norms for the three deontic concepts have been studied by Broersen [2003] and Ju and van Eijck [2019]. The modalities we will introduce below have different logical properties from those introduced in the latter works – but more on this in Section 6.4.2. In addition, neither Broersen [2003] nor Ju and van Eijck [2019] include a notion of sub-ideality in their systems. Second, Sergot and Craven [2006], Craven and Sergot [2008], and Kulicki and Trypuz [2017] supply labeled transition systems with a deontic classification of both actions and states. Besides differing from our framework in important details concerning the interaction between "green" states and transitions, these proposals do not cover norms about sequences of actions. A notion of sub-ideality appears in Sergot and Craven [2006] (but not in Craven and Sergot [2008] and Kulicki and Trypuz [2017]), where the authors rank transitions according to how well they satisfy a given system of norms. There are many interesting connections between the latter proposal and our own. Yet, Sergot and Craven [2006] design a formalism for *defining* labeled transition systems rather than using labeled transition systems to *interpret* a modal language. A full comparison is thus beyond the scope of the present chapter.

6.2 Framing the system

Our proposal is based on the idea that, in order to address the issues arising from the paradoxes of $\mathrm{PD}_e\mathrm{L}$, we need to be able to draw more distinctions than those that can be drawn in that system. Let us introduce them with a simple example.

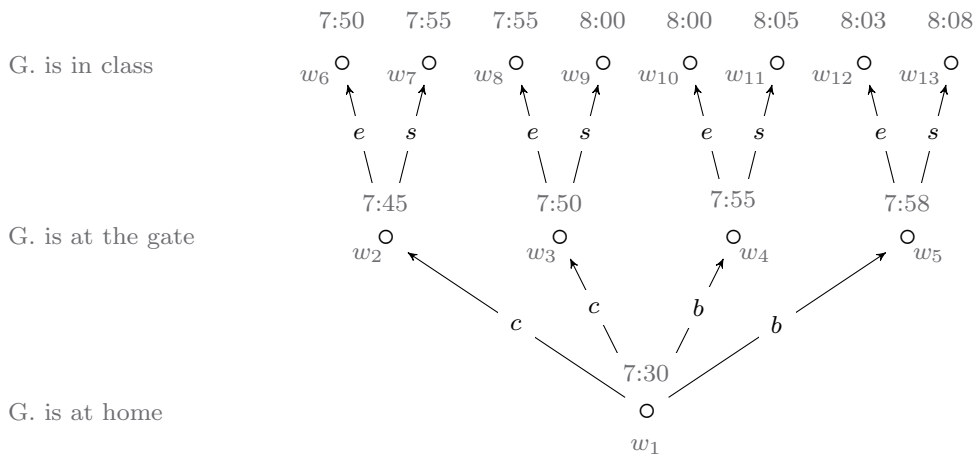


Figure 6.1: A transition system representing Example 6.2.1

6.2.1. Example. In most Italian cities, school starts at 8:00am, although kids are recommended to be in their classrooms at least 5 minutes before. Suppose that it is 7:30am and a kid, Giacomo, can go to school either by car, with his aunt, or by bike. If he goes by car, he will be at the school gate at 7:45am if his aunt exceeds the speed limit and at 7:50am if his aunt respects all traffic regulations. If he goes by bike, he will be at the school gate at 7:55am if everything goes smoothly and at 7:58am if he accidentally falls off his bike. Once he is at the school gate, Giacomo needs 5 minutes to get to his classroom if he takes the elevator and 10 minutes if he takes the stairs.

Leaving deontic elements aside, Example 6.2.1 can be represented as the transition system depicted in Figure 6.1, where circles stand for possible states, arrows for action tokens (thought of as transitions between states, cf. Chapter 2.3.2) and arrow-labels for the action types instantiated by the corresponding labeled tokens. The following action types are represented: *going to school by car* (c), *going to school by bike* (b), *taking the elevator* (e), and *taking the stairs* (s). The propositions on the left of the diagram are true at the states on their right and the time above a state indicates that it is that time at that state. So, at the bottom-most state w_1 it is 7:30am and Giacomo is at home; at w_2 it is 7:45am and Giacomo is at the school gate; and so on. The two c-transitions starting at w_1 indicate that there are two ways in which action type c can be executed at it: one that results in Giacomo's being at the school gate at 7:45am (his aunt does not respect the speed limit) and one that results in Giacomo's being at the school gate at 7:50am (his aunt respects all traffic regulations). Similarly, the two b-transitions

starting at w_1 indicate that there are two ways of executing action type b at it, one resulting in Giacomo's being at the school gate at 7:55am (he does not fall off his bike) and one resulting in Giacomo's being at the school gate at 7:58am (he falls off his bike).

Notice that only *one-step actions* are depicted in Figure 6.1. This is not a coincidence: In our formalism below, propositions about sequences of actions will be expressed by modal formulas that only involve one-step actions. Sequences of actions will thus *not* appear among the terms of our language.[13] In addition, we will take on Meyer's [1988] basic representation of actions and *not* include any explicit reference to agents, leaving the study of a proper multi-agent extension of our proposal to future work. In the semantics, we will represent the information encoded in a labeled transition system like the one in Figure 6.1 by using the following ingredients: (1) a relation R_1 relating every state w to the states that can be reached from w in one step (called *directly accessible from* w), (2) a function f_{dn} determining which actions have *just been done* at any state w, (3) a relation R_{ac} relating every state w to the states that can be reached from w in one or more steps (called *accessible from* w). If w' is directly accessible from w (i.e., $wR_{[1]}w'$), we will label the transition (w, w') with the action types that have just been done at w'.

Given a set of norms, the actions and courses of action that can be performed at a state w can be classified in terms of the states that are (directly) accessible from w. In Example 6.2.1, the relevant norms are the norm that kids ought to be in their classrooms by 8:00am (preferably by 7:55am) and the norm that drivers ought to respect the speed limit. Here is then a first deontic category of states:

- States at which all relevant norms are satisfied are *ideal*.

Figure 6.2 represents the same transition system as Figure 6.1 where the states inside the ellipses are ideal (more on the black states in a moment). In our frames, ideal states will be modeled in the standard way by means of a set **ideal** of ideal states. We will make the following assumption concerning ideal states:

Assumption 1 *For any state w, there is an ideal state that is accessible from w.*

According to Assumption 1, it is possible to recover, eventually, from any breach of the law. In the present chapter, we will thus ignore both inconsistent systems of norms, whose prescriptions cannot possibly be satisfied, and situations of deontic tragedy characterized by a persistent state of violation.

So far so good. But, as we saw in Section 6.1, ideality is often too high a standard for guiding actions: in many cases, we find ourselves in "sad" or,

[13] This will allow us to assume a simple algebra of actions, without worrying too much about the behavior of the operator of action negation, whose interaction with the operator of sequential composition is far from trivial [see, e.g., Broersen, 2004; Dignum and Meyer, 1990; Ju and van Eijck, 2019; Wansing, 2004].

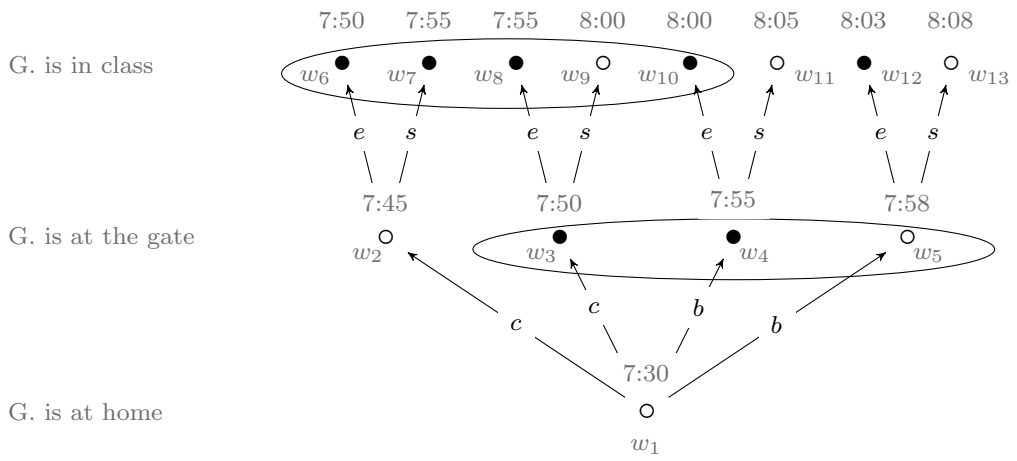


Figure 6.2: A deontic transition system representing Example 6.2.1

as we will say, *substandard* circumstances. There are two senses in which the circumstances can be substandard:

1. *backward-looking*: some norm has just been violated;

2. *forward-looking*: it is not possible to act in accordance with all norms.[14]

"Substandard" in the backward looking sense is the same as "not ideal," so let us call "*not ideal*" the states that satisfy the description in item 1 and reserve "*substandard*" for the states that satisfy the description in item 2 (we will say that a state is *standard* when it is not substandard). The two senses are independent of one another. In our example, the state at which Giacomo is at the school gate at 7:40am is not ideal (the traffic regulations have just been violated) but standard (the kid can act in accordance with with all relevant norms). On the other hand, the state at which Giacomo is at the school gate at 7:58am is substandard (there is nothing the kid can do to be in his classroom by 8:00am), but ideal (the kid has done nothing wrong – accidents can happen!). It is at substandard states like the latter that ideality is not an appropriate principle of action guidance. As Hansson [1969] or Lewis [1974] would put it, at such states one ought to do *the best* they can. The problem in front of us is to bring together the two principles "comply with the laws" and "do the best you can."

Let us start by introducing a second deontic category of states:

[14]This kind of situation generates what we will call *local normative conflict* in the next chapter.

- States that are directly accessible from a state w by doing the best that can be done at w are *optimal relative to* w.

By "doing the best that can be done at w" we mean two things: (1) complying with the given norms in the best possible way if w is standard (i.e., the given norms can be complied with at w), and (2) violating the given norms in the least bad way possible if w is substandard (i.e., the given norms cannot be complied with at w). To illustrate, look again at Figure 6.2. The black circles represent optimal states relative to the states that directly access them. So, for instance, state w_8 (where Giacomo is in his classroom at 7:55am) is optimal relative to state w_2 (where Giacomo is at the school gate at 7:50am). The reason is that, by being in his classroom 5 minutes before school starts, the kid complies with the given norms in the best possible way. At the same time, state w_{12} (where Giacomo is in his classroom at 8:03) is optimal relative to w_5 (where the kid is at the school gate at 7:58am). The reason is that, by being in his classroom as soon after 8:00am as he can, the kid violates the given norms in the least bad way possible. This notion of *relative optimality* will be modeled in our frames by means of a relation $R_{[op]}$ relating every state w with the states that are optimal relative to w. Given how we read "optimal state relative to w," the following assumptions are uncontroversial:

Assumption 2 *For any state w, there is an optimal state relative to w.*

Assumption 3 *If w is standard, then only ideal states are optimal relative to w.*

The notions of ideality and relative optimality allow us to distinguish, for any state w, four types of states directly accessible from w:

1. *green!*: ideal and optimal relative to w (none if w is substandard);
2. *green*: ideal, possibly optimal relative to w (none if w is substandard);
3. *orange*: optimal relative to w but not ideal (none if w is standard);
4. *red*: neither green nor orange.

This coloring of states makes it easy to track different kinds non-compliant behavior: unlike violations of the relevant norms occurring at red states, violations of the relevant norms occurring at orange states are, in an obvious sense, excusable. In addition, the one-step actions that can be performed at a state w (we will call them the action types *executable at* w from now on) can now be classified in terms of the colors of their possible outcomes at w in the following way:

1. *green!*: some outcomes at w are green! (none if w is substandard);
2. *green*: some outcomes at w are green (none if w is substandard);

- **Ideal state**: all norms are complied with.
- **Not ideal state**: some norm are violated.
- **Standard state**: some executable action leads to an ideal state.
- **Substandard state**: no executable action leads to an ideal state.
- **Optimal state relative to** w: state that is directly accessible from w and
  (a) ideal in the best possible way if w is standard, and
  (b) not ideal in the least bad possible way if w is substandard.

Table 6.2: Terminology applying to possible states

3. *orange*: some outcomes at w are orange (none if w is standard);
4. *red*: all outcomes at w are red.

With our running example, the action of going to school by car (c) and the action of going to school by bike (b) are both green! (hence green) at w_1; the action of taking the elevator (e) is green! (hence green) at all states in middle row, except for w_5, where it is orange; finally, the action of taking the stairs (s) is green! (hence green) at w_2, green but not green! at w_3, and red at w_4 and w_5.

With these distinctions in place, an obvious principle of action guidance is

For any state w, perform either a green or an orange action at w.

As we will see in more details in Section 6.4.1, the suggested principle intuitively says "try to comply with the law if you can and, otherwise, try to do the best you can." It thus combines the maxims "comply with the law" and "do the best you can" in a natural way. We will use the new maxim to define and study notions of (process-) permission, prohibition, and obligation after introducing our formal system. For future reference, Table 6.2 summarizes the key terminology introduced so far.

6.3 The dynamic deontic logic $\mathsf{PD}_e\mathsf{LO}$

In this section, we present the dynamic deontic logic with optimality $\mathsf{PD}_e\mathsf{LO}$. After introducing its syntax and semantics, we provide a sound and complete axiomatization and flag some facts concerning its relation with $\mathrm{PD}_e\mathrm{L}$.

6.3.1 Syntax and semantics

As the language of PDL and $\mathrm{PD}_e\mathrm{L}$, the language $\mathcal{L}_{\mathsf{PD}_e\mathsf{LO}}$ of the dynamic deontic logic $\mathsf{PD}_e\mathsf{LO}$ contains two categories of expressions: (names of) action types and

formulas. We start by fixing a countable non-empty set *Atm* of (names of) atomic action types and a countable non-empty set *Prop* of propositional variables.

6.3.1. DEFINITION (Syntax of $\mathcal{L}_{\mathsf{PD}_e\mathsf{LO}}$). Let *Atm* and *Prop* be defined as above. The set *Types* of (names of) action types of $\mathcal{L}_{\mathsf{PD}_e\mathsf{LO}}$ is generated by the following grammar:

$$\alpha := a \,|\, \overline{\alpha} \,|\, \alpha \cup \beta \,|\, \alpha \cap \beta$$

where $a \in Atm$. We use lower case letters from the beginning of the alphabet a, b, c for elements of *Atm* and greek letters from the beginning of the alphabet α, β, γ for elements of *Types*. The set of formulas of $\mathcal{L}_{\mathsf{PD}_e\mathsf{LO}}$, also denoted with $\mathcal{L}_{\mathsf{PD}_e\mathsf{LO}}$, is generated by the following grammar:

$$\varphi := p \,|\, id \,|\, dn(\alpha) \,|\, \neg\varphi \,|\, (\varphi \wedge \varphi) \,|\, \mathsf{A}\varphi \,|\, [1]\varphi \,|\, [ac]\varphi \,|\, [op]\varphi$$

where $p \in Prop$ and $\alpha \in Types$. The abbreviations for the other Boolean connectives are standard. In addition, we use $\mathsf{E}\varphi$, $\langle 1\rangle\,\varphi$, $\langle ac\rangle\,\varphi$, and $\langle op\rangle\,\varphi$ as abbreviations for $\neg\mathsf{A}\neg\varphi$, $\neg[1]\neg\varphi$, $\neg[ac]\neg\varphi$, $\neg[op]\neg\varphi$ respectively. Parentheses are eliminated according to the usual conventions.

We think of *Types* as a set of *one-step actions* and give to the action combinators $\overline{\cdot}$, $\cup$, and $\cap$ the intuitive interpretation presented in Section 6.1.2. We assume that an action token can instantiate different atomic action types.

The formulas of $\mathcal{L}_{\mathsf{PD}_e\mathsf{LO}}$ can be divided into *ontic formulas* and *deontic formulas*. The ontic formulas are built from the action predicate dn, the universal modality A, and the modalities for accessibility [1] and $[ac]$. A formula like $dn(\alpha)$ says that an action of type α has just been performed, while a formula like $\mathsf{A}\varphi$ says (as usual) that φ is true at all possible states. As for the other ontic modalities: $[1]\varphi$ means "φ will necessarily be true in one step, no matter which actions are executed now" and is true at a state w when φ is true at all states that are *directly accessible* from w; $[ac]\varphi$ means "φ will necessarily be true in the future, no matter which actions will be executed from now on" and is true at a state w when φ is true at all states that are *accessible* from w. The deontic formulas of $\mathcal{L}_{\mathsf{PD}_e\mathsf{LO}}$ are built form the propositional constant id, which intuitively says that the present state is ideal, and the modality for optimality $[op]$. A formula like $[op]\varphi$ says that realizing φ is optimal and is true at a state w when φ is true at all states that are optimal relative to w.

Dynamic modalities in PDL-style can be introduced in $\mathcal{L}_{\mathsf{PD}_e\mathsf{LO}}$ by using the following abbreviations:

6.3.2. DEFINITION. Where $\alpha \in Types$ and $\varphi \in \mathcal{L}_{\mathsf{PD}_e\mathsf{LO}}$, $[\alpha]\varphi := [1](dn(\alpha) \to \varphi)$. The dual modality $\langle\alpha\rangle$ is defined in the standard way as $\neg[\alpha]\neg$.

So, $[\alpha]\varphi$ says that φ will necessarily be true in one step if an action of type α is performed.

The semantics for $\mathcal{L}_{\mathsf{PD}_e\mathsf{LO}}$ is based on the notion of $\mathsf{PD}_e\mathsf{LO}$ frame. $\mathsf{PD}_e\mathsf{LO}$ frames are basically transitions systems supplied with two deontic components, namely a set of ideal states and a relation modeling relative optimality.

6.3.3. Definition ($\mathsf{PD}_e\mathsf{LO}$ frame). A $\mathsf{PD}_e\mathsf{LO}$ frame is a tuple

$$\langle W, R_{[1]}, R_{[ac]}, f_{dn}, R_{[op]}, \mathbf{ideal} \rangle$$

where $W \neq \varnothing$ is a set of possible states, $R_{[1]} \subseteq W \times W$ is the *direct accessibility* relation, $R_{[ac]} \subseteq W \times W$ is the *accessibility* relation, $f_{dn} : Types \to 2^W$ assigns to every action type the set of states where a token of that type has just been performed, $R_{[op]} \subseteq W \times W$ is the relation of *relative optimality*, and $\mathbf{ideal} \subseteq W$ is a set of *ideal states*. The elements of $\mathsf{PD}_e\mathsf{LO}$ frames satisfy the following conditions:

1. *Properties of $R_{[1]}$ and $R_{[ac]}$*: for all $w, w_1, w_2, w_3 \in W$,
   Seriality of $R_{[1]}$: there is $w' \in W$ such that $wR_{[1]}w'$.
   Accessibility of directly accessible states: if $w_1R_{[1]}w_2$, then $w_1R_{[ac]}w_2$.
   Transitivity of $R_{[ac]}$: if $w_1R_{[ac]}w_2$ and $w_2R_{[ac]}w_3$, then $w_1R_{[ac]}w_3$.
2. *Properties of f_{dn}*: for all $\alpha, \beta \in Types$,
   Union: $f_{dn}(\alpha \cup \beta) = f_{dn}(\alpha) \cup f_{dn}(\beta)$.
   Intersection: $f_{dn}(\alpha \cap \beta) = f_{dn}(\alpha) \cap f_{dn}(\beta)$.
   Complement: $f_{dn}(\overline{\alpha}) = W \setminus f_{dn}(\alpha)$.
3. *Properties of $R_{[op]}$ and* **ideal**: for all $w, w_1, w_2 \in W$,
   Accessibility of an ideal state: for some $w' \in W$, $wR_{[ac]}w'$ and $w' \in \mathbf{ideal}$.
   Direct accessibility of optimal states: if $w_1R_{[op]}w_2$, then $w_1R_{[1]}w_2$.
   Seriality of $R_{[op]}$: there is $w' \in W$ such that $wR_{[op]}w'$.
   Conditional ideality of optimal states: if there is $w' \in W$ such that $wR_{[1]}w'$ and $w' \in \mathbf{ideal}$, then, for all w' such that $wR_{[op]}w'$, $w' \in \mathbf{ideal}$.

For any relation $R \subseteq W \times W$ and $w \in W$, we define $R(w) = \{w' \in W \mid wRw'\}$.

The conditions on $\mathsf{PD}_e\mathsf{LO}$ frames ensure that their components behave as described in Section 6.2. The conditions on the "ontic components" $R_{[ac]}$, $R_{[1]}$, and f_{dn} are obvious: they guarantee that something can always be done (seriality of $R_{[1]}$), that states that are directly accessible are accessible (accessibility of directly accessible states), that the states reached by performing a sequence of actions at an accessible state are themselves accessible (transitivity of $R_{[ac]}$), and, finally, that the action operators behave like Boolean operators (properties of f_{dn}). The conditions on the "deontic components" $R_{[op]}$ and **ideal** codify the definitions and assumptions discussed in Section 6.2. The first condition corresponds to

Assumption 1 that, for any state w, there is an ideal state accessible from w. The second condition reflects the fact that we defined "optimal state relative to w" as "*state that is directly accessible from* w by doing the best one can at w." The last two conditions state, respectively, that, for any state w, there is an optimal state relative to w (Assumption 2), and that, if w is standard, then only ideal states are optimal relative to w (Assumption 3) – where a state w is a $\mathsf{PD}_e\mathsf{LO}$ frame is *standard* when $R_{[1]}(w) \cap \mathbf{ideal} \neq \varnothing$ and *substandard* otherwise.

The notion of $\mathsf{PD}_e\mathsf{LO}$ model and the notion truth are defined as follows.

6.3.4. Definition ($\mathsf{PD}_e\mathsf{LO}$ model). A $\mathsf{PD}_e\mathsf{LO}$ model is a tuple $\langle F, \nu \rangle$, where F is a $\mathsf{PD}_e\mathsf{LO}$ frame and $\nu : Prop \to 2^W$ is a valuation function.

6.3.5. Definition (Semantics for $\mathcal{L}_{\mathsf{PD}_e\mathsf{LO}}$). Given a $\mathsf{PD}_e\mathsf{LO}$ model M, truth of a formula $\varphi \in \mathcal{L}_{\mathsf{PD}_e\mathsf{LO}}$ at a state w in M, denoted $M, w \models \varphi$, is defined recursively. Truth of atomic propositions and the Boolean connectives is defined as usual. The remaining cases are as follows:

$$
\begin{array}{lll}
M, w \models id & \text{iff} & w \in \mathbf{ideal} \\
M, w \models dn(\alpha) & \text{iff} & w \in f_{dn}(\alpha) \\
M, w \models \mathsf{A}\varphi & \text{iff} & \text{for all } w' \in W,\ M, w' \models \varphi \\
M, w \models [1]\varphi & \text{iff} & \text{for all } w' \in W, \text{ if } wR_{[1]}w', \text{ then } M, w' \models \varphi \\
M, w \models [ac]\varphi & \text{iff} & \text{for all } w' \in W, \text{ if } wR_{[ac]}w', \text{ then } M, w' \models \varphi \\
M, w \models [op]\varphi & \text{iff} & \text{for all } w' \in W, \text{ if } wR_{[op]}w', \text{ then } M, w' \models \varphi
\end{array}
$$

For any formula $\varphi \in \mathcal{L}_{\mathsf{PD}_e\mathsf{LO}}$, $[\![\varphi]\!]^M = \{w \in W \mid M, w \models \varphi\}$ is the truth-set of φ in M. Reference to the model M is omitted when it is clear from the context.

Given any $\mathsf{PD}_e\mathsf{LO}$ model M and action type $\alpha \in Types$, let us define a relation $R_\alpha \subseteq W \times W$ by setting, for all $w, w' \in W$,

$$wR_\alpha w' \text{ iff } wR_{[1]}w' \text{ and } w' \in f_{dn}(\alpha).$$

Intuitively, among all transitions determined by $R_{[1]}$, R_α selects the α-transitions. The defined modality $[\alpha]\varphi$ has then the following semantics, in line with the standard semantics for dynamic modalities:

$$
\begin{array}{lll}
M, w \models [\alpha]\varphi & \text{iff} & M, w \models [1](dn(\alpha) \to \varphi) \\
& \text{iff} & \text{for all } w' \in W, \text{ if } wR_{[1]}w' \text{ and } w' \in f_{dn}(\alpha), \text{ then } M, w' \models \varphi \\
& \text{iff} & \text{for all } w' \in W, \text{ if } wR_\alpha w', \text{ then } M, w' \models \varphi
\end{array}
$$

We will use dynamic modalities repeatedly in Section 6.4.

6.3.2 Axiomatization

The axiom system $\mathsf{PD}_e\mathsf{LO}$ is defined by the axioms and rules in Table 6.3. The items in the first five rows at the top are standard. Axioms in groups (I) and (II) express the constraints on the ontic components of $\mathsf{PD}_e\mathsf{LO}$ frames: the axioms for *dn* reflect the fact that action types are interpreted over a Boolean algebra of sets; the inclusion axioms express that A is interpreted over the set of all possible states ($\mathsf{Inc1}$), that direct accessibility implies accessibility ($\mathsf{Inc2}$), and that relative optimality implies direct accessibility ($\mathsf{Inc3}$). The remaining axioms correspond to our three assumptions on ideality and relative optimality: axiom D for $[op]$ says that some optimal state of affairs can be realized (Assumption 2), $\mathsf{Ax4}$ that, eventually, an ideal state can be realized (Assumption 1), and $\mathsf{Ax5}$ that, if an ideal state can be realized in one step, then only what is ideal is also optimal.

(CPL)	Classical propositional tautologies	(MP)	From φ and $\varphi \to \psi$, infer ψ
($\mathsf{S5_A}$)	The axiom schemas of $\mathsf{S5}$ for A	($\mathsf{RN_A}$)	From φ, infer $\mathsf{A}\varphi$
($\mathsf{S4}_{[ac]}$)	The axiom schemas of $\mathsf{S4}$ for $[ac]$		
($\mathsf{K}_{[1]}$)	$[1](\varphi \to \psi) \to ([1]\varphi \to [1]\psi)$		
($\mathsf{KD}_{[op]}$)	The axiom schemas of KD for $[op]$		
	(I) Axioms for *dn*		**(II) Inclusion axioms**
($\mathsf{Ax1}$)	$dn(\overline{\alpha}) \leftrightarrow \neg dn(\alpha)$	($\mathsf{Inc1}$)	$\mathsf{A}\varphi \to [ac]\varphi$
($\mathsf{Ax2}$)	$dn(\alpha \cup \beta) \leftrightarrow dn(\alpha) \vee dn(\beta)$	($\mathsf{Inc2}$)	$[ac]\varphi \to [1]\varphi$
($\mathsf{Ax3}$)	$dn(\alpha \cap \beta) \leftrightarrow dn(\alpha) \wedge dn(\beta)$	($\mathsf{Inc3}$)	$[1]\varphi \to [op]\varphi$
	(III) Axioms for ideal		
($\mathsf{Ax4}$)	$\langle ac \rangle id$	($\mathsf{Ax5}$)	$\langle 1 \rangle id \to [op]id$

Table 6.3: The axiom system $\mathsf{PD}_e\mathsf{LO}$

The following theorem, which is a consequence of Definition 6.3.2, the axioms in group (I) and axiom $\mathsf{K}_{[1]}$, shows that our system is powerful enough to interpret the fragment of $\mathrm{PD}_e\mathrm{L}$ without the action operator of sequential composition. The proof is straightforward and is thus omitted.

6.3.6. Theorem. *The following are theorems of* $\mathsf{PD}_e\mathsf{LO}$*:*

1. $[\alpha](\varphi \to \psi) \to ([\alpha]\varphi \to [\alpha]\psi)$
2. $[\overline{\alpha}]\varphi \leftrightarrow [\alpha]\varphi$
3. $[\alpha]\varphi \vee [\beta]\varphi \to [\alpha \cap \beta]\varphi$
4. $[\alpha \cup \beta]\varphi \leftrightarrow [\alpha]\varphi \wedge [\beta]\varphi$
5. $[\overline{\alpha \cap \beta}]\varphi \leftrightarrow [\overline{\alpha}]\varphi \wedge [\overline{\beta}]\varphi$
6. $[\overline{\alpha}]\varphi \vee [\overline{\beta}]\varphi \to [\overline{\alpha \cup \beta}]\varphi$

To conclude this section, it is not difficult to prove the following result.

6.3.7. THEOREM. *The axiom system* $\mathsf{PD}_e\mathsf{LO}$ *is sound and strongly complete with respect to the class of all* $\mathsf{PD}_e\mathsf{LO}$ *frames.*

The proof of soundness is a matter of routine validity check. The proof of completeness is based on the construction of a canonical model for $\mathsf{PD}_e\mathsf{LO}$ and, by paying attention to the universal modality, proceeds in an entirely standard way [see Blackburn et al., 2001, Chapter 4.2; Goranko and Passy, 1992]. We only provide the definition of the canonical model for $\mathsf{PD}_e\mathsf{LO}$ and leave the details of the proof to the reader.

Let $\mathcal{W}$ be the set of all maximal consistent sets of $\mathsf{PD}_e\mathsf{LO}$. Where $w \in \mathcal{W}$ and $\blacksquare \in \{[1], [ac], [op]\}$, define: $w/\blacksquare = \{\varphi \in \mathcal{L}_{\mathsf{PD}_e\mathsf{LO}} \mid \blacksquare\varphi \in w\}$.

6.3.8. DEFINITION (Canonical $\mathsf{PD}_e\mathsf{LO}$ model for w_0). The canonical $\mathsf{PD}_e\mathsf{LO}$ model for $w_0 \in \mathcal{W}$ is a tuple $M^c = \left\langle W^c, R^c_{[1]}, R^c_{[ac]}, f^c_{dn}, R^c_{[op]}, \mathbf{ideal}^c, \nu^c \right\rangle$, where

- $W^c = \{w \in \mathcal{W} \mid w_0/\mathsf{A} \subseteq w\}$;
- $R^c_{[1]} \subseteq W^c \times W^c$ is such that, for all $w, w' \in W^c$, $wR^c_{[1]}w'$ iff $w/[1] \subseteq w'$;
- $R^c_{[ac]} \subseteq W^c \times W^c$ is such that, for all $w, w' \in W^c$, $wR^c_{[ac]}w'$ iff $w/[ac] \subseteq w'$;
- $f_{dn} : Types \to 2^{W^c}$ is such that, for all $\alpha \in Types$ and $w \in W^c$, $w \in f^c_{dn}(\alpha)$ iff $dn(\alpha) \in w$;
- $R^c_{[op]} \subseteq W^c \times W^c$ is such that, for all $w, w' \in W^c$, $wR^c_{[op]}w'$ iff $w/[op] \subseteq w'$;
- $\mathbf{ideal}^c \subseteq W^c$ is such that, for all $w \in W^c$, $w \in \mathbf{ideal}^c$ iff $id \in w$;
- $\nu^c : Prop \to 2^{W^c}$ is such that, for all $w \in W^c$, $w \in \nu^c(p)$ iff $p \in w$.

Given Definition 6.3.8, the proof of the Truth Lemma (for every $w \in W^c$ and $\varphi \in \mathcal{L}_{\mathsf{PD}_e\mathsf{LO}}$, $M^c, w \models \varphi$ iff $\varphi \in w$) and the proof that M^c is a $\mathsf{PD}_e\mathsf{LO}$ model proceed in the usual way. This is sufficient to conclude that every consistent set $\Gamma \subseteq \mathcal{L}_{\mathsf{PD}_e\mathsf{LO}}$ is satisfiable in a $\mathsf{PD}_e\mathsf{LO}$ model.

6.4 Deontic operators and paradoxes

We are now ready to address the issues we discussed at the end of Section 6.1.2. We start from studying prescriptions concerning one-step actions. After introducing appropriate deontic operators for actual prescriptions, we will turn to an analysis of prescriptions concerning sequences of actions.

6.4.1 From ideal to actual prescriptions

Let us begin by going back to the deontic classification of states and actions discussed in Section 6.2. Since $\mathsf{PD}_e\mathsf{LO}$ frames incorporate a distinction between ideal and relatively optimal states, the notions of green!, green, orange, and red states relative to a given state w can be modeled in a straightforward way:

6.4.1. Definition. Where $\langle W, R_{[1]}, R_{[ac]}, f_{dn}, R_{[op]}, \mathbf{ideal}\rangle$ is a $\mathsf{PD}_e\mathsf{LO}$ frame and $w \in W$,

$$\begin{array}{lcllcl} gr!(w) & = & R_{[op]}(w) \cap \mathbf{ideal} & or(w) & = & R_{[op]}(w) \setminus \mathbf{ideal} \\ gr(w) & = & R_{[1]}(w) \cap \mathbf{ideal} & rd(w) & = & R_{[1]}(w) \setminus (gr(w) \cup or(w)) \end{array}$$

Observe that $gr(w) \subseteq R_{[1]}(w)$ and $red(w) \subseteq R_{[1]}(w)$ by definition and $gr!(w) \subseteq R_{[1]}(w)$ and $or(w) \subseteq R_{[1]}(w)$ by the condition of direct accessibility of optimal states. So, Definition 6.4.1 provides a classification of the states that are *directly accessible* from a given state. Specifically, green! states are both ideal and optimal relative to that state; green states are ideal but possibly not optimal relative to it; orange states are optimal relative to it but not ideal; finally, red states are neither ideal nor optimal relative to it. In line with the terminology introduced above, we will say that a state w in a $\mathsf{PD}_e\mathsf{LO}$ frame is *substandard* if $gr(w) = \varnothing$, i.e., if no ideal state is directly accessible from w (we will say that w is standard otherwise). The next simple propositions and corollary will be useful later on:

6.4.2. Proposition. *For any state w in a $\mathsf{PD}_e\mathsf{LO}$ frame, (a) if $gr(w) \neq \varnothing$, then $or(w) = \varnothing$ and (b) if $gr(w) = \varnothing$, then $gr!(w) = \varnothing$.*

Proof:
(a) If $R_{[1]}(w) \cap \mathbf{ideal} \neq \varnothing$, then $R_{[op]}(w) \subseteq \mathbf{ideal}$ by the condition of conditional ideality of optimal states. Hence, $or(w) = R_{[op]}(w) \setminus \mathbf{ideal} = \varnothing$. (b) By the condition of direct accessibility of optimal states, $R_{[op]}(w) \subseteq R_{[1]}(w)$. Hence, if $R_{[1]}(w) \cap \mathbf{ideal} = \varnothing$, then $gr!(w) = R_{[op]}(w) \cap \mathbf{ideal} = \varnothing$. □

6.4.3. Corollary. *For any state w in a $\mathsf{PD}_e\mathsf{LO}$ frame, $R_{[op]}(w) = gr!(w) \cup or(w)$.*

Proof:
It is obvious from Def. 6.4.1 that $gr!(w) \cup or(w) \subseteq R_{[op]}(w)$. For the other direction, if $R_{[1]}(w) \cap \mathbf{ideal} \neq \varnothing$, then (1) $gr!(w) \cup or(w) = gr!(w)$ by Prop. 6.4.2 (a) and (2) $R_{[op]}(w) \subseteq \mathbf{ideal}$ by the condition of conditional ideality of optimal states. By (2), $R_{[op]}(w) = \mathbf{ideal} \cap R_{[op]}(w) = gr!(w)$, and so $R_{[op]}(w) = gr!(w) \cup or(w)$ by (1). If $R_{[1]}(w) \cap \mathbf{ideal} = \varnothing$, then (3) $gr!(w) \cup or(w) = or(w)$ by Prop. 6.4.2 (b). In addition, (4) $R_{[op]}(w) = R_{[op]}(w) \setminus \mathbf{ideal} = or(w)$, as $R_{[op]}(w) \subseteq R_{[1]}(w)$ by the condition of

direct accessibility of optimal states. Hence, $R_{[op]}(w) = gr!(w) \cup or(w)$ by (3) and (4). □

Together with Proposition 6.4.2, Corollary 6.4.3 is a reformulation of the idea that the relation $R_{[op]}$ selects the best states given the circumstances, i.e., green! states where the law is complied with in the best possible way if the circumstances are standard and orange states where the law is violated in the least bad way possible if the circumstances are substandard. The defined coloring of states induces a coloring of the actions executable at a given state. For any state w, let $exe(w) = \{\alpha \in Types \mid R_\alpha(w) \neq \varnothing\}$ be the set of actions *executable* at w.

6.4.4. Definition. Where w is a state in a $\mathsf{PD}_e\mathsf{LO}$ frame and $\alpha \in exe(w)$,

$$\alpha \in \underline{gr!}(w) \text{ iff } f_{dn}(\alpha) \cap gr!(w) \neq \varnothing \qquad \alpha \in \underline{or}(w) \text{ iff } f_{dn}(\alpha) \cap or(w) \neq \varnothing$$
$$\alpha \in \underline{gr}(w) \text{ iff } f_{dn}(\alpha) \cap gr(w) \neq \varnothing \qquad \alpha \in \underline{rd}(w) \text{ iff } \alpha \in exe(w) \setminus (\underline{gr}(w) \cup \underline{or}(w))$$

So an action is green! (resp. green or orange) at w if some of its possible outcomes at w are green! (resp. green or orange), and it is red at w if all of its possible outcomes are red at w.

The introduced classifications of states and actions can be used to interpret deontic operators codifying the two maxims "comply with the law" and "do the best you can." The first maxim can be understood either in a global sense ("realize those states of affairs that obtain at all ideal states") or in a local sense ("realize those states of affairs that obtain at all directly accessible ideal states") and it gives rise to *ideal prescriptions*. The second maxim gives rise to what we will call *optimal prescriptions*.

6.4.5. Definition (Standard deontic operators). Where $\varphi \in \mathcal{L}_{\mathsf{PD}_e\mathsf{LO}}$,

	Obligation	**Permission**	**Prohibition**
Ideal - global	$O^{\mathsf{A}}\varphi := \mathsf{A}(id \to \varphi)$	$P^{\mathsf{A}}\varphi := \neg O^{\mathsf{A}}\neg\varphi$	$F^{\mathsf{A}}\varphi := O^{\mathsf{A}}\neg\varphi$
Ideal - local	$O^{1}\varphi := [1](id \to \varphi)$	$P^{1}\varphi := \neg O^{1}\neg\varphi$	$F^{1}\varphi := O^{1}\neg\varphi$
Optimal	$O^{op}\varphi := [opt]\varphi$	$P^{op}\varphi := \neg O^{op}\neg\varphi$	$F^{op}\varphi := O^{op}\neg\varphi$

Let D be any of the deontic operators from the above table. For any $\alpha \in Types$, we set: $D\alpha := Ddn(\alpha)$.

The evaluation rules for the deontic operators from Definition 6.4.5 are displayed in Table 6.4.[15] The operators O^{A}, P^{A}, F^{A} are standard deontic operators in the tradition of Anderson [1958] and Kanger [1957], while the operators O^1, P^1, F^1 are – modulo the action operator of sequential composition – generalizations

[15] The evaluation rules for prohibition operators are analogous to those for obligation operators and are thus omitted.

Obligation	Permission
$M, w \models O^{\mathsf{A}}\varphi$	$M, w \models P^{\mathsf{A}}\varphi$
iff $\mathbf{ideal} \subseteq [\![\varphi]\!]$	iff $\mathbf{ideal} \cap [\![\varphi]\!] \neq \varnothing$
$M, w \models O^{1}\varphi$	$M, w \models P^{1}\varphi$
iff $R_{[1]}(w) \cap \mathbf{ideal} \subseteq [\![\varphi]\!]$	iff $R_{[1]}(w) \cap \mathbf{ideal} \cap [\![\varphi]\!] \neq \varnothing$
iff $gr(w) \subseteq [\![\varphi]\!]$	iff $gr(w) \cap [\![\varphi]\!] \neq \varnothing$
$M, w \models O^{op}\varphi$	$M, w \models P^{op}\varphi$
iff $R_{[op]}(w) \subseteq [\![\varphi]\!]$	iff $R_{[op]}(w) \cap [\![\varphi]\!] \neq \varnothing$
iff $(gr!(w) \cup or(w)) \subseteq [\![\varphi]\!]$	iff $(gr!(w) \cup or(w)) \cap [\![\varphi]\!] \neq \varnothing$

Table 6.4: Semantics for O^{A}, P^{A}, O^{1}, P^{1}, O^{op}, P^{op}

of the deontic operators of PD$_e$L [cf. Section 6.1.2].[16] The operators in the last group roughly correspond to the unconditional deontic operators from Hansson [1969] and Lewis [1974]: they pick out the best states relative to certain circumstances. According to the evaluation rules in Table 6.4, ideal local prescriptions are determined by *green states*, while optimal prescriptions are determined by *green! states* if the circumstances are standard and by *orange states* otherwise [cf. Proposition 6.4.2]. When applied to action types, P^1 and P^{op} can be used to express the color of an action. In fact, for any state w in any $\mathsf{PD}_e\mathsf{LO}$ model M and any $\alpha \in Types$, we have:

1. $\alpha \in \underline{gr!}(w)$ iff $gr!(w) \cap f_{dn}(\alpha) \neq \varnothing$ iff $R_{[op]}(w) \cap f_{dn}(\alpha) \neq \varnothing$ and $gr(w) \neq \varnothing$ iff $M, w \models P^{op}\alpha \wedge \langle 1 \rangle\, id$;
2. $\alpha \in \underline{gr}(w)$ iff $gr(w) \cap f_{dn}(\alpha) \neq \varnothing$ iff $M, w \models P^{1}\alpha$;
3. $\alpha \in \underline{or}(w)$ iff $or(w) \cap f_{dn}(\alpha) \neq \varnothing$ iff $R_{[op]}(w) \cap f_{dn}(\alpha) \neq \varnothing$ and $gr(w) = \varnothing$ iff $M, w \models P^{op}\alpha \wedge [1]\neg id$;
4. $\alpha \in \underline{rd}(w)$ iff $\alpha \in exe(w) \setminus (\underline{gr}(w) \cup \underline{or}(w))$ iff $M, w \models \langle \alpha \rangle \top \wedge \neg P^{1}\alpha \wedge \neg P^{op}\alpha$.

In order to illustrate the use of the notions of ideal and optimal prescriptions in the analysis of concrete cases, let us go back to Example 6.2.1 as represented in Figure 6.2. Let φ_1 stand for the proposition "Giacomo just entered his class and it is later than 8:00am" (so, φ_1 is true at w_{11}, w_{12}, and w_{13}) and φ_2 for the proposition "a car ride has just been completed and the speed limit has been violated" (so, φ_2 is true at w_2). Recall that c, b, e, and s stand for the action

[16] Definitions 6.3.2 and 6.4.5 and the axioms on dn ensure that the formulas $P^1\alpha \leftrightarrow \langle \alpha \rangle\, id$, $F^1\alpha \leftrightarrow \neg P^1\alpha$, and $O^1\alpha \leftrightarrow \neg P^1\overline{\alpha}$ are theorems of $\mathsf{PD}_e\mathsf{LO}$.

types *going to school by car*, *going to school by bike*, *taking the elevator*, and *taking the stairs*. Then,

- $O^{\mathsf{A}}\neg\varphi_1$ and $O^{\mathsf{A}}\neg\varphi_2$ are true at all states. The two formulas can be used to express (roughly) that, according to the given norms, the kid ought to be in his class by 8:00am and the speed limit ought to be respected when driving. Observe that the operators O^{A} and P^{A} are not suitable to describe how the kid is supposed to comply with the given norms in specific situations: $P^{\mathsf{A}}s$ is true at all states even if at w_4 the kid would surely be late for class if he took the stairs.

- $O^1 e$ is true at w_4. The formulas expresses that, given the circumstances, the kid has to take the elevator to comply with the norms. Notice that the operators O^1 and P^1 are not suitable to describe what the kid is supposed to do in substandard situations: $O^1\,\bot$ (a trivial obligation) is true at w_5 even if, intuitively, at this state the kid ought to go to his classroom as quickly as possible.

- $O^{op}e$ is true at w_5. The formula expresses that, given the (substandard) circumstances, Giacomo ought to take the elevator to do his best. Observe that the operators O^{op} and P^{op} are not suitable to describe what the kid is supposed to do to comply with the norms in standard circumstances: $\neg P^{op}s$ is true at w_3 even if, intuitively, at this state the kid would not do anything wrong by taking the stairs. At this state, the prohibition $\neg P^{op}s$ expresses what is preferable rather than what the law requires.

So, the standard operators do capture important types of prescriptions. Yet, none of them is suitable, by itself, to model what *actually* may or ought to be done, given the circumstances: ideal permissions and obligations (global or local) are pointless in substandard situations, while optimal obligations are too demanding in standard situations. What we need are deontic operators that adapt to the circumstances by picking green states if the circumstances are standard and orange states if they are not – that is, operators that encode the maxim "comply with the law if you can; otherwise, do the best you can." Consider an immediate consequence of Proposition 6.4.2:

6.4.6. Fact. For any state w in a $\mathsf{PD}_e\mathsf{LO}$ frame,

1. if $gr(w)\neq\varnothing$, then $gr(w)\cup or(w)=gr(w)$ (hence, $\underline{gr}(w)\cup\underline{or}(w)=\underline{gr}(w)$);
2. if $gr(w)=\varnothing$, then $gr(w)\cup or(w)=or(w)$ (hence, $\underline{gr}(w)\cup\underline{or}(w)=\underline{or}(w)$).

Fact 6.4.6 suggests that a suitable evaluation rule for an operator O^a for *actual obligation* could be the following: $M,w\models O^a\varphi$ iff $gr(w)\cup or(w)\subseteq[\![\varphi]\!]$. It is immediately seen that, thus interpreted, formulas like $O^a\varphi$ can be introduced as

An unlabeled arrow from formula A to formula C means that $A \to C$ is valid in the class of $\mathsf{PD}_e\mathsf{LO}$ models. A labeled arrow from A to C means that $A \to C$ is valid in the class of $\mathsf{PD}_e\mathsf{LO}$ models, *provided that* A is strengthened with additional conditions: $\boxed{!_1}$ means $\langle 1 \rangle id$ and $\boxed{!_2}$ means $\neg \langle 1 \rangle id$.

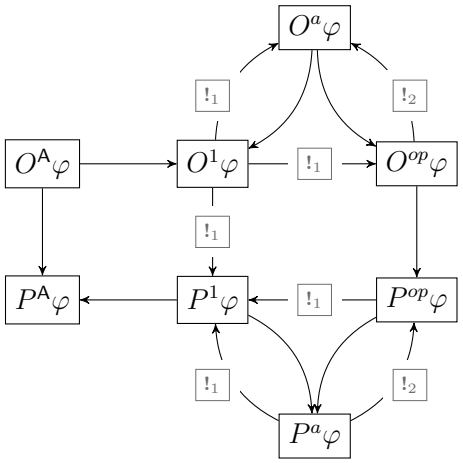


Figure 6.3: Logical relations between deontic operators

abbreviations in our language. In fact, the following holds for any state w in any $\mathsf{PD}_e\mathsf{LO}$ model M:[17]

$$\begin{aligned} M, w \models O^a\varphi \quad & \text{iff} \quad gr(w) \cup or(w) \subseteq [\![\varphi]\!] \\ & \text{iff} \quad gr(w) \subseteq [\![\varphi]\!] \text{ and } gr!(w) \subseteq [\![\varphi]\!] \text{ and } or(w) \subseteq [\![\varphi]\!] \\ & \text{iff} \quad gr(w) \subseteq [\![\varphi]\!] \text{ and } gr!(w) \cup or(w) \subseteq [\![\varphi]\!] \\ & \text{iff} \quad M, w \models O^1\varphi \wedge O^{op}\varphi \end{aligned}$$

This leads us to the following definition.

6.4.7. Definition (Actual deontic operators). Where $\varphi \in \mathcal{L}_{\mathsf{PD}_e\mathsf{LO}}$,

$$O^a\varphi := O^1\varphi \wedge O^{op}\varphi \qquad P^a\varphi := \neg O^a \neg \varphi \qquad F^1\varphi := O^a\neg\varphi$$

Where D^a is any of the three defined operators and $\alpha \in Types$, we set: $D^a\alpha := D^a dn(\alpha)$.

The main logical relations between the standard and actual deontic operators are summarized in Figure 6.3. Some facts are worth noticing. First, $O^{op}\varphi$ does not entail $O^a\varphi$. The reason is that, at a standard state w, some ideal states directly accessible from w may not be optimal relative to w (take, e.g., state w_9 relative to w_3 in Figure 6.2). This suggests that our notion of actual obligation is in line with those consequentialist theories according to which one ought to realize an acceptable amount of positive consequences rather than a maximum amount thereof [see, e.g., Portmore, 2011]. Second, neither $O^{\mathsf{A}}\varphi$ nor $O^1\varphi$ entail

[17] For the second line, recall that $gr!(w) \subseteq gr(w)$.

$O^a\varphi$. The reason is that, unlike ideal obligations, actual obligations can always be satisfied in one step ($O^a\varphi$ entails $\langle 1\rangle\,\varphi$). This is, in fact, the key difference between ideal and actual obligations. Finally, let us mention that our operators are subject to two well-know deontic paradoxes:

Ross's paradox. [Ross, 1944] $O^a\varphi$ entails $O^a(\varphi \vee \psi)$. So, if it is obligatory to bake a cake, it is obligatory to either bake a cake or throw the ingredients away.

Paradox of free-choice permission. [von Wright, 1968] $P^a(\varphi \vee \psi)$ does not entail $P^a\varphi \wedge P^a\psi$, even if, intuitively, if it is permitted to choose between two options, then both options should be permitted. What is worse, since $P^a\varphi$ entails $P^a(\varphi \vee \psi)$, adding a "free-choice principle" to the effect that $P^a(\varphi \vee \psi)$ entails $P^a\varphi \wedge P^a\psi$ would lead to the undesirable consequence that everything is permitted if something is.

McCarty [1983], Meyer [1988] and Segerberg [1982] have independently suggested that, when the resources of dynamic logic are available, the aforementioned paradoxes could be overcome by introducing strong notions of permission and obligation. According to these proposals, an action is *strongly permitted (obligatory)* just in case it is permitted (obligatory) in the ideal local sense and all ways of doing it lead to an ideal state. In the present framework, this idea can be implemented by introducing operators P^s and O^s with the following semantics:

1. $M, w \models P^s\varphi$ iff $w \in [\![P^a\varphi]\!]$ and $R_{[1]}(w) \cap [\![\varphi]\!] \subseteq gr(w) \cup or(w)$.

   So: $M, w \models P^s dn(\alpha)$ iff $w \in [\![P^a\alpha]\!]$ and $R_{[1]}(w) \cap f_{dn}(\alpha) \subseteq gr(w) \cup or(w)$ iff $w \in [\![P^a\alpha]\!]$ and $R_\alpha(w) \subseteq gr(w) \cup or(w)$

2. $M, w \models O^s\varphi$ iff $w \in [\![O^a\varphi]\!]$ and $R_{[1]}(w) \cap [\![\varphi]\!] \subseteq gr(w) \cup or(w)$.

   So: $M, w \models O^s dn(\alpha)$ iff $w \in [\![O^a\alpha]\!]$ and $R_{[1]}(w) \cap f_{dn}(\alpha) \subseteq gr(w) \cup or(w)$ iff $w \in [\![O^a\alpha]\!]$ and $R_\alpha(w) \subseteq gr(w) \cup or(w)$.

Yet, such notions of permission and obligation are well-known to be too strong: no action that might result in a violation of the law because of a mistake, an accident, or because of the presence of normative constraints can be permitted or obligatory in this sense. We think that defining dedicated notions of *choice-permission* and *choice-obligation* could be a more promising solution. Specifically, we could introduce formulas like $P^a(\varphi + \psi)$ as abbreviations for $P^a\varphi \wedge P^a\psi$ and formulas like $O^a(\varphi + \psi)$ as abbreviations for $O^a(\varphi \vee \psi) \wedge P^a(\varphi + \psi)$. The defined expressions would capture several intuitions: that a choice is permitted only if both alternatives are permitted; that being permitted to do something does not imply that one should not be careful about how they do that thing; finally, that a choice is required when (a) not realizing either alternative is prohibited and (b) choosing between them is permitted. An interesting question, which we leave to future research, is whether one could arrive at the proposed definitions by

introducing a new connective + such that $\varphi + \psi$ is the choice of realizing either φ or ψ. We conjecture that this may be achieved by importing techniques from state based semantics [see, e.g., Aloni, 2007] or inquisitive/truthmaker semantics [see, e.g., Aloni, 2018; Aloni and Ciardelli, 2013; Anglberger et al., 2016, Fine, 2018a; Fine, 2018b].

6.4.2 Process norms

The operators for actual prescriptions introduced by means of Definition 6.4.7 apply to one-step actions. Our present aim is to use them to express process norms that are sensitive to whether the circumstances along the process are standard or not. Let us start from permission. Semantically, the idea is simple [see also van der Meyden, 1996]:

* *Performing an action of type α and then performing an action of type β is permitted if there is at least one "good" execution of α that is followed by a "good" execution of β.*

We have learned in the previous section that, in order to obtain operators that are sensitive to the circumstances, "good" should be read as "ending in an ideal state" if the circumstances are standard and as "ending in an optimal state" if the circumstances are substandard. This suggests that a formula like $P^a(\alpha;\beta)$, saying that it is permitted to do α and then β, should have the following semantics: $M, w \models P^a(\alpha;\beta)$ iff there is a state $w' \in (gr(w) \cup or(w)) \cap f_{dn}(\alpha)$ such that $\beta \in \underline{gr}(w') \cup \underline{or}(w')$. It is easy to see that, given this interpretation, $P^a(\alpha;\beta)$ is definable in our language: where w is any state from any $\mathsf{PD}_e\mathsf{LO}$ model M,

$$\begin{array}{rll} M, w \models P^a(\alpha;\beta) & \text{iff} & \text{there is } w' \in (gr(w) \cup or(w)) \cap f_{dn}(\alpha) \\ & & \text{such that } (gr(w') \cup or(w')) \cap f_{dn}(\beta) \neq \varnothing \\ & \text{iff} & \text{there is } w' \in (gr(w) \cup or(w)) \cap f_{dn}(\alpha) \\ & & \text{such that } w' \in [\![P^a\beta]\!] \\ & \text{iff} & (gr(w) \cup or(w)) \cap [\![dn(\alpha) \wedge P^a\beta]\!] \neq \varnothing \\ & \text{iff} & M, w \models P^a(dn(\alpha) \wedge P^a\beta) \end{array}$$

By generalizing to any finite sequence of action types, we obtain the following definition.

6.4.8. Definition (Process-permission/prohibition). Where $\alpha_1, \ldots, \alpha_n \in Types$,

$$P^a(\alpha_1;\ldots;\alpha_n) := P^a(dn(\alpha_1) \wedge P^a(dn(\alpha_2) \wedge P^a(\ldots P^a(dn(\alpha_{n-1}) \wedge P^a\alpha_n)\ldots))$$

In addition, $F^a(\alpha_1;\ldots;\alpha_n)$ abbreviates $\neg P^a(\alpha_1;\ldots;\alpha_n)$.

A formula like $F^a(\alpha;\beta)$ thus means that either α is red or β is red at all states reached by a "good" execution of α. The following proposition follows immediately from Definition 6.4.8, Definition 6.4.7, and axiom Inc3.

6.4.9. Proposition. *Where $\alpha_1,\ldots,\alpha_n \in Types$, the following is a theorem of* $\mathsf{PD}_e\mathsf{LO}$*:*

$$P^a(\alpha_1;\ldots;\alpha_n) \to P^a\alpha_1 \wedge \langle\alpha_1\rangle\, P^a(\alpha_2;\ldots;\alpha_n)$$

Accordingly, a sequence of actions is permitted only if the first action in the sequence is permitted and there is a way to perform it after which it is permitted to complete the sequence. The next example, adapted from Craven and Sergot [2008], will help us highlighting some additional key properties of the proposed notions of process- permission and prohibition.

6.4.10. Example. Suppose that there is a flat consisting of two rooms, the left room and the right room, connected by a single doorway. A girl, Alice, and two boys, Bob and Carl, move around the flat. Only one person can move through the doorway at a time. In addition, when Alice, Bob, and Carl are all in the same room, Alice has a way to prevent the two boys from leaving that room. There are four rules in the flat: rule n_1 states that no girl may be in a room alone with only one boy; rule n_2 states that, in case n_1 is violated, girls in the left room ought to move to the right room; rule n_3 states that, in case n_1 is violated, boys in the left room ought to stay where they are; finally, rule n_4 states that, in case n_1 is violated, boys in the right room ought to move to the left room. Currently, Alice is in the left room with Bob, while Carl is in the right room.

Although a full analysis of Example 6.4.10 requires a multi-agent framework, we can consider what Alice may and may not do by treating Bob and Carl as "Nature." Figure 6.4 illustrates how we can represent the example in a $\mathsf{PD}_e\mathsf{LO}$ model by taking Alice's perspective (only a few relevant states are depicted). Labels above states are mnemonics for the position of the three agents in the flat. For instance, AB|C means that Alice and Bob are in the left room and Carl is in the right room and ABC|$\varnothing$ that the three agents are all in the left room. The arrow-labels represent the following action types: *staying* (s), *moving to a different room* (m), *preventing the boys from leaving a room* (b). Since Bob and Carl are treated as "Nature," it is intended that each action type is instantiated when Alice performs an action of that type. Note that, in line with the story, the action type *preventing the boys from leaving a room* is only executable at w_2, where Alice, Bob, and Carl are all in the same room. As before, states in the ellipses are ideal and black states are optimal relative to the states from which they are directly accessible.[18]

[18] In identifying optimal states, we have assumed that a state w' is optimal relative to w when at all other states that are directly accessible from w there are at least as many violations as

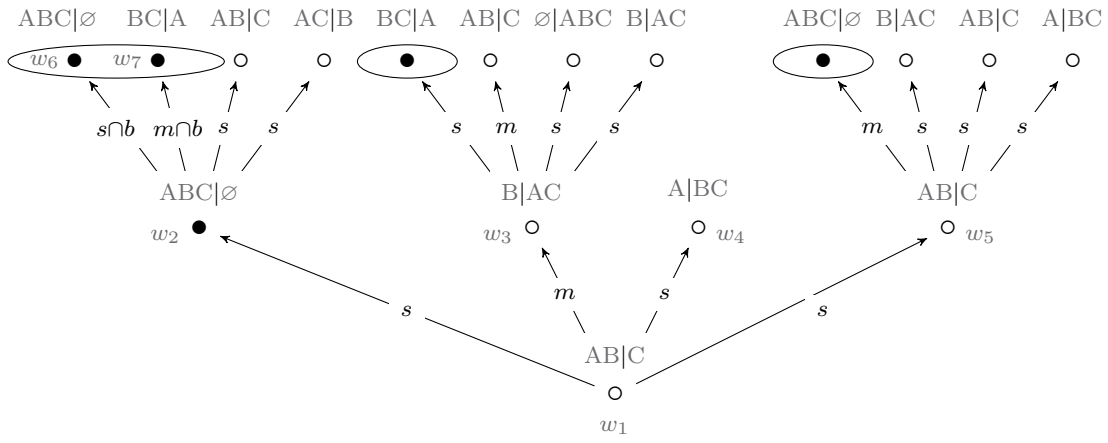


Figure 6.4: A way to model Example 6.4.10 in $\mathsf{PD}_e\mathsf{LO}$

The example is interesting for several reasons. First, although w_1 is both substandard and not ideal, there is a one-step action that Alice may perform at this state, namely *staying*. In fact, since there is an s-transition from w_1 to w_2 and w_2 is optimal relative to w_1, s is orange at w_1, and so $P^a(s)$ is true at this state. Even more, there is a sequence of actions that Alice may perform at w_1, namely *staying and then moving to a different room*. In fact, since m is green at the end-state of the "good" s-transition from w_1 to w_2, $P^a(s;m)$ is true at w_1. The latter fact shows that our notion of process-permission effectively adapts to the circumstances as they arise during a course of actions: it picks orange states as substandard states are reached and green states as standard states are reached.

Second, $\langle m\rangle\, P^a(s)$ is true at w_1 because there is an m-transition from w_1 to w_3 and s is orange at w_3. Yet, $P^a(m)$ is false at w_1 because no m-transition starting at w_1 leads to an optimal state. Hence, $P^a(m;s)$ is false at w_1 by Proposition 6.4.9.[19] So, even if it is possible that, after moving to a different room, Alice may stay there, the process *moving to a different room and then staying* is prohibited at w_1 because *moving to a different room* is prohibited in the first place. This shows that the proposed notion of process-permission is subject neither to van der Meyden's paradox ($\langle\alpha\rangle\, P^a\beta$ does not entail $P^a(\alpha;\beta)$) nor to Angleberger's paradox (since $F^a(m) \wedge \langle m\rangle\, P^a(s)$ is true at w_1, $F^a\alpha$ does not entail $[\alpha]F^a\beta$), as

at w' [cf. Sergot and Craven, 2006]. For instance, at the states directly accessible from w_1 the following rules have just been violated: n_2 at w_2; n_1 and n_4 at w_3; n_2, n_3, n_4 at w_4; and n_1, n_2, n_4 at w_5. Since the fewest number of violations occur at w_2, w_2 is the only optimal state relative to w_1.

[19] Recall that $\mathsf{PD}_e\mathsf{LO}$ is sound with respect to the class of $\mathsf{PD}_e\mathsf{LO}$ models, so the implication in Proposition 6.4.9 is valid on this class.

it was desired.

Finally, it is worth noticing that $P^a(s;s)$ is true at w_1 (for the same reason why $P^a(s;m)$ is) and, yet, $[s]P^a(s)$ is not. In fact, there is an s-transition from w_1 to w_5 and s is red at w_5. So, although at w_1 Alice is permitted to stay in the right room and then keep staying there, it is possible that, after staying in the right room, Alice ends up in a state where she is not permitted to stay there anymore. After staying in the right room, Alice will be permitted to keep staying there only if things go in the best possible way.

This shows that, under the present proposal, process-permissions do not set permissions step-by-step once and for all. Rather, they provide the agent with an indication of what would be, in principle, the best course of action to take, without excluding that things might change in due course. Other notions of process-permission advanced in the literature [cf., e.g., Broersen, 2003; Ju and van Eijck, 2019; van der Meyden, 1996] are different in this respect, as they satisfy the following principle:

$$P^a(\alpha_1;\ldots;\alpha_n) \to P^a\alpha_1 \wedge [\alpha_1]P^a(\alpha_2;\ldots;\alpha_n) \tag{P1}$$

We think that permission operators satisfying P1 are too strong, in the same sense in which the operator of strong permission discussed at the end of Section 6.4.1 is. With a simple example, suppose that there is a limit of 1,000€ on daily cash withdrawal. Then, we may well be permitted to withdraw cash twice in a day, even if, in case we withdraw 1,000€ the first time, we will not be allowed to withdraw cash a second time.

Turning to process-obligations, let us start by introducing a bit of terminology. A *path* in a $\mathsf{PD}_e\mathsf{LO}$ model is any sequence of states $w_1w_2\ldots w_n$ such that $w_iR_{[1]}w_{i+1}$, for all $i < n$. We say that a path $w_1w_2\ldots w_n$ is *good* when $w_{i+1} \in gr(w_i) \cup or(w_i)$ for all $i < n$. Here is then an intuitive way of thinking of process-obligations:

- *It is obligatory to perform an action of type α and then an action of type β if performing α and then β is necessary to be on a good path.*

Formally, this coincides with giving the following semantics to formulas like $O^a(\alpha;\beta)$ (read: "it is obligatory to do α and then β"):

$$M, w \models O^a(\alpha;\beta) \text{ iff } gr(w) \cup or(w) \subseteq f_{dn}(\alpha) \text{ and, for all } w' \in gr(w) \cup or(w), \\ gr(w') \cup or(w') \subseteq f_{dn}(\beta)$$

It follows immediately from the evaluation rule of $O^a\varphi$ [see p. 170] that, thus interpreted, $O^a(\alpha;\beta)$ can be introduced as an abbreviation of $O^a\alpha \wedge O^aO^a\beta$. Definition 6.4.11 generalizes this idea to any finite sequence of actions types.

6.4.11. Definition (Process-obligation). Where $\alpha_1,\ldots,\alpha_n \in Types$,

$$O^a(\alpha_1;\ldots;\alpha_n) := O^a\alpha_1 \wedge O^aO^a\alpha_2 \wedge \cdots \wedge \underbrace{O^a\ldots O^a}_{n}\alpha_n$$

To illustrate, let us go back to Figure 6.4. $O^a(s)$ is true at w_1, since the only orange state relative to w_1 is w_2 and the transition linking w_1 and w_2 is an s-transition. $O^aO^a(b)$ is also true at w_1, since the only green states relative to w_2 (which is the only orange state relative to w_1) are w_6 and w_7 and the transitions linking w_2 to these states are b-transitions. Hence, $O^a(s;b)$ is true at w_1: Alice ought to stay in the left room and then prevent the boys from leaving it.

On the other hand, notice that $[s]O^a(b)$ is not true at w_1. In fact, there is an s-transition from w_1 to w_5 and the action type b is not executable at w_5. Since, as we observed above, $O^a(b)$ entails $\langle b\rangle\top$, it follows that $O^a(b)$ is false at w_5. Hence, even if at w_1 Alice ought to stay in the left room and then prevent the boys from leaving it, it is possible that, after staying in the left room, it is in fact not obligatory to prevent the boys from leaving it: as in the case of process-permission, the latter obligation is triggered only if everything goes in the best possible way in the first step. Again, this distinguishes our notion of process-obligation from other notions proposed in the literature [cf., e.g., Broersen, 2003; Ju and van Eijck, 2019; van der Meyden, 1996], which satisfy the principle:

$$O(\alpha_1;\ldots;\alpha_n)\to O\alpha_1\wedge[\alpha_1]O(\alpha_2;\ldots;\alpha_n) \tag{P2}$$

As before, we think that obligation operators satisfying P2 are too strong. For instance, even if we ought to register our change of address at the Municipality and then keep the records for at least three years, we ought not to keep the record for at least three years if we misspell the address (in fact, in this case, we ought to correct the registered address as soon as possible and throw the wrong record away). Still, it follows immediately from Definition 6.4.11, Definition 6.4.7, and axiom Inc3 that the right-to-left direction of P2 is a theorem of $\mathsf{PD}_e\mathsf{LO}$ (and intuitively so):

6.4.12. Proposition. *Where $\alpha_1,\ldots,\alpha_n\in Types$, the following is a theorem of $\mathsf{PD}_e\mathsf{LO}$:*

$$O^a\alpha_1\wedge[\alpha_1]O^a(\alpha_2;\ldots;\alpha_n)\to O^a(\alpha_1;\ldots;\alpha_n)$$

Definition 6.4.11 and Proposition 6.4.12 ensure that we can implement Meyer's [1988] formalization of the Chisholm's set in $\mathsf{PD}_e\mathsf{LO}$, by maintaining the distinction between compliant-with-duty and contrary-to-duty obligations that appears in his system. Specifically, following Meyer [1988], we can formalize the Chisholm's set as follows (recall that ρ and κ stand for the action types *robbing* and *calling the police*):

$$\text{(a) } O^a(\overline{\rho}),\ \text{(b) } [\overline{\rho}]O^a(\overline{\kappa}),\ \text{(c) } [\rho]O^a(\kappa)$$

By Proposition 6.4.12, (a) and (b) entail $O^a(\overline{\rho};\overline{\kappa})$. In addition, since $O^a(\overline{\rho})$ entails $\neg O^a(\rho)$, (a) entails $\neg O^a(\rho;\kappa)$ by Definition 6.4.11. So, while compliant-with-duty obligations direct to the performance of a "good" course of action, contrary-to-duty obligations do not. Finally, the $\mathsf{PD}_e\mathsf{LO}$ model in Figure 6.5 shows that a

(1) Since there is an a-transition from w_1 to w_2 and w_2 is the only optimal state relative to w_1, $O^a(a)$ is true at w_1. (2) The only a-transition from w_1 ends at w_2, where the b-transition to w_4 leads to the only optimal state relative to w_2. So, $[a]O^a(b)$ is true at w_1. (3) Since the only $\overline{a}$-transition from w_1 ends at w_3, where the only transition to an optimal state is a b-transition, $[\overline{a}]O^a(b)$ is true and $[\overline{a}]O^a(\overline{b})$ is false at w_1.

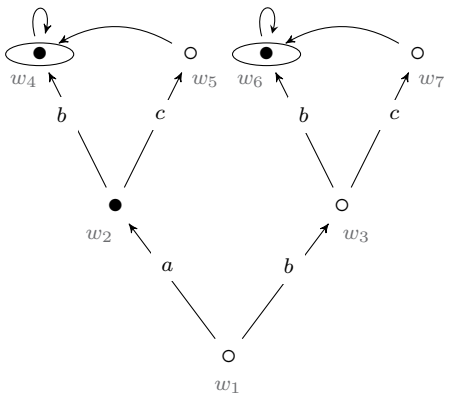


Figure 6.5: A $\mathsf{PD}_e\mathsf{LO}$ countermodel for $O^a(\alpha) \wedge [\alpha]O^a(\beta) \rightarrow [\overline{\alpha}]O^a(\overline{\beta})$.

primary obligation of form $O^a(\alpha)$ together with a compliant-with-duty obligation of form $[\alpha]O^a(\beta)$ does not entail a CTD obligation of form $[\overline{\alpha}]O^a(\overline{\beta})$. Hence, the difficulty affecting Meyer's proposal arising from Angleberger's paradox is overcome.

6.5 Conclusion

At the beginning of this chapter we saw that there is a tension between two key principles of action guidance in deontic logic. The principle "comply with the law" supports the introduction of deontic operators interpreted in terms of *ideality*, like the operators of SDL, of systems of alethic deontic logic [Anderson, 1958; Kanger, 1957], and of $\mathrm{PD}_e\mathrm{L}$ [Meyer, 1988]. The principle "do the best you can" supports the introduction of deontic operators interpreted in terms of *optimality*, like the operators of dyadic deontic systems in the tradition of Hansson [1969] and Lewis [1974]. In Sections 6.2 and 6.4.1, we claimed that, under this interpretation, neither family of operators is adequate to capture *actual prescriptions*, expressing what we may or ought to do in concrete situations: Ideal deontic operators based on the former principle are unsatisfactory in substandard situations in which we cannot comply with the law. Optimal operators based on the latter principle are unsatisfactory in standard situations in which we can.

In tandem with the paradoxes affecting $\mathrm{PD}_e\mathrm{L}$, this motivated us to design a dynamic deontic system, $\mathsf{PD}_e\mathsf{LO}$, incorporating both a notion of ideality and a notion of optimality. The system is based on few basic semantic ingredients: a distinction between accessible and directly accessible states at the ontic level and a distinction between ideality and relative optimality at the deontic level. These key distinctions provided the basis for the formulation of a rich deontic

classification ("coloring") of states and actions, which, as shown by the discussion of Examples 6.2.1 and 6.4.10, allowed us to meet several *desiderata* for an analysis of norm-governed transition systems: (1) the possibility of distinguishing different categories of non-compliant behavior (i.e., performing an orange action *vs* performing a red action); (2) the possibility of keeping track of prescriptions at different normative levels (i.e., global ideal, local ideal, optimal, actual); (3) the possibility of effectively identifying, in a way that is both sensitive to whether the circumstances are (sub)standard and compatible with accidents and mistakes, which (courses of) actions actually may or ought to be performed; finally, (4) the possibility of distinguishing compliant-with-duty, process norms from contrary-to-duty, goal norms. The key for the latter two results was the introduction of *actual deontic operators* based on the idea of merging the two aforementioned principles of action guidance in the maxim "try to comply with the law if you can; otherwise, do the best you can."

We observed more than once in the course of the chapter that our system lacks the resources to properly represent multi-agent scenarios. Although we can determine what an individual agent may or ought to do by treating the other agents as "Nature" (as we did in the analysis of Example 6.4.10), we cannot study how the prescriptions applying to the different individual agents interact with one another, nor can we determine whose fault it is if a substandard state is reached or some norm is violated. In the next chapter, we begin to address this issue by developing a multi-agent deontic logic in which the behavior of different agents is possibly regulated by different sets of norms. We will focus on a particular kind of interaction between such sets of norms, i.e., normative conflicts, and on a particular kind of substandard situations, i.e., those resulting from the presence of a normative conflict.

Chapter 7

Normative conflicts in a dynamic logic of norms and codes

In the last decades, the study of normative conflicts in deontic logic has been guided by two main issues: first, developing *conflict tolerant deontic systems*, typically by tweaking the logical principles of SDL [see Goble, 2005, 2009, 2013]; second, given a conflict tolerant system, developing *solution procedures* for deontic conflicts based on, e.g., priority relations between obligations [see, e.g., Hansen, 2006, 2014; Horty, 2003, 2007, 2012] or selection functions in the tradition of input/output logics [Makinson and van der Torre, 2000, 2001; Parent and van der Torre, 2013]. In the present chapter, we supplement these lines of research by proposing a framework to model the *dynamics that gives rise to a conflict*. We think that assuming this perspective contributes to two main lines of research: (1) The study of solution procedures, since conflicts generated in different ways might require diverse solutions. (2) The study of responsibility in substandard situations in which one or more agents cannot avoid violating a norm [see Chapter 6], since how a conflict arises is crucial to determine who (if anybody) should take responsibility for the violations that inevitably result from it.

Our guiding idea is that, in order to model the origin of a conflict, we need a way to explicitly refer to the *normative codes* guiding the behavior of the agents involved. We think of an agent's normative code as the set of norms that the agent accepts as binding. By *adding new norms* to their normative codes, agents can then generate conflicts either within their own codes or between their codes and the codes of other agents. In this view, the interaction between different kinds of normative sources, specifically norms and codes, plays an essential role. This motivates us to design a logic with the resources to explicitly represent the following elements: norms and key relations between them; agents' codes and their interaction with norms; the dynamics corresponding to the inclusion of a new norm in an agent's code. The resulting deontic logic of norms and codes is characterized by two main features. First, it is *explicit* in the sense of explicit modal logics [Artemov, 2008; Fitting, 2005], as norms and codes are

introduced in it as explicit sources of prescriptions. Second, it is *dynamic* in the sense of Dynamic Epistemic Logic (DEL) [Baltag et al., 1998; van Benthem, 2011; van Ditmarsch et al., 2008], as it represents procedures of code update that correspond, in the semantics, to model transformations.

Before we start, let us emphasize once more that, unlike other logics of norms, like input/output logics [see Parent and van der Torre, 2013] or deontic systems based on default logics [see Horty, 2012], our system is *not* devised for representing logical relations between norms nor for deriving normative solutions given a conflict between the norms in a code. Rather, it is specifically built to help us analyzing the *genesis* of such conflicts. For the reader familiar with applications of DEL to deontic logic, it is also worth mentioning that the update procedure we propose is different from deontic variants of preference update [see, e.g., van Benthem et al., 2014; van Benthem and Liu, 2014; Liu, 2011, Chapter 11; Yamada, 2008]. It is, on the other hand, similar to operations of "considering" studied in the context of awareness logic [see, e.g., Velázquez-Quesada, 2009; van Benthem and Velázquez-Quesada, 2010] and operations of "becoming aware" studied in the context of dynamic evidence logic [see, e.g., Baltag et al., 2012, 2014]. The difference between the latter approaches and our own lies in the specific way the "explicit" and "intensional" components interact both in the static logic and in the definition of the update procedure.

Outline. Sections 7.1 and 7.2 gradually introduce the static logic of norms and codes NC as an extension of a logic of norms that we call N. The two sections are organized in the same way: we first introduce the syntax and semantics of the logic under consideration [Sections 7.1.1 and 7.2.1] and then provide a sound and complete axiomatization [Sections 7.1.2 and 7.2.2]. In Section 7.3 we make the system NC dynamic by defining a procedure of code update, corresponding to the event that an agent accepts a new norm in her code. This gives rise to the dynamic logic of norms and codes DNC. In the first half of Section 7.4 [Section 7.4.1], we illustrate how DNC can be used to keep track of the source of a conflict by analyzing the paradigmatic examples of Antigone and Gandhi. In the second half of the section [Section 7.4.2], we show that some key features of cases of civil disobedience and conscientious objection can be captured in a simple refinement of DNC (called DNC^{+}). Section 7.5 concludes by pointing to possible developments for future works.

This chapter is based on Canavotto and Giordani [2018].

7.1 The logic of norms N

The logic of norms N is based on the idea that norms are elementary normative sources that agents can adopt as directives for acting.[1] We assume a basic conception of norms, according to which every norm is associated with a set of *explicit prescriptions* from which *implicit prescriptions* can be derived.

7.1.1. Example. Article 20 of the Doctorate Regulations of the University of Amsterdam is a norm that *explicitly* prescribes the following: (1) "In addition to the supervisor (and co-supervisor), the Doctorate Committee consist of at least five and at most seven remaining members," (3) "The voting members of the Committee shall consists of a majority of full professors," (5) "At least half of the voting Committee members must be affiliated with the University." Given (1), (3), and (5), Article 20 *implicitly* prescribes that, if the Doctorate Committee consists of three full professors affiliated with the University of Amsterdam and two associate professors not affiliated with it, then the remaining member of the Committee ought to be a full professor and may be either affiliated with the University of Amsterdam or not.

In the following, we will call the set of explicit prescriptions of a norm its *explicit content* and the set of its implicit prescriptions its *implicit content*. We will assume that every norm is essentially characterized by its explicit content, i.e., that its explicit content does not change from a situation to another. In addition, we will take norms to be *consistent*, in the sense that no contradiction can be derived from their explicit contents. This means that the explicit prescriptions of every norm can be jointly fulfilled, or *satisfied*, at some possible state and that inconsistencies can only be found between the prescriptions of different norms.

Besides these conditions, which are specific of norms, we will assume that all normative sources (norms or codes) obey two additional minimal requirements. First, they explicitly prescribe their own satisfaction: insofar as it is a directive, every normative source at least directs us to its fulfillment. Second, when a normative source prescribes that a norm ought to be satisfied, it also prescribes everything that that norm explicitly prescribes. So, prescribing the satisfaction of a norm amounts to including its explicit content. With these preliminaries in place, let us now introduce the syntax and semantics of the logic N.

7.1.1 Syntax and semantics

Let us start by fixing a non-empty countable set $Prop$ of propositional variables and a non-empty countable set $Norms$ of (names of) norms (we use n possibly with superscripts $n', n'', \ldots$ for elements of the latter set).

[1] In this chapter, "agent" refers to any entity that has the ability to assume certain norms as binding. Hence, agents include persons, artificial agents, communities, organizations, states, and so on.

7.1.2. Definition (Syntax of $\mathcal{L}_{\mathsf{N}}$). Let $Norms$ and $Prop$ be defined as above. The set of formulas of $\mathcal{L}_{\mathsf{N}}$, also denoted with $\mathcal{L}_{\mathsf{N}}$, is generated by the following grammar:

$$\varphi := p \mid sat(n) \mid n : \varphi \mid \neg\varphi \mid (\varphi \wedge \varphi) \mid \mathsf{A}\varphi \mid [ac]\varphi$$

where $p \in Prop$ and $n \in Norms$. The abbreviations for the other Boolean connectives are defined as usual. We use $\mathsf{E}\varphi$ and $\langle ac \rangle$ as abbreviations for $\neg\mathsf{A}\neg\varphi$ and $\neg[ac]\neg\varphi$ respectively. We will adopt the usual rules for the elimination of parentheses.

$\mathcal{L}_{\mathsf{N}}$ includes both ontic and deontic formulas. The ontic formulas $\mathsf{A}\varphi$ and $[ac]\varphi$ have the same intended meaning as in Chapter 6: $\mathsf{A}\varphi$ says that φ holds at all possible states and $[ac]\varphi$ says that φ is true at all (directly) accessible states. The two modalities are intended to capture the distinction between what is logically possible and what is possible in specific circumstances. The deontic formulas of $\mathcal{L}_{\mathsf{N}}$ include $sat(n)$, which means "norm n is satisfied," and $n : \varphi$, which means "norm n explicitly prescribes φ."

The interaction between ontic and deontic modalities allows us to express the presence of two different kinds of conflicts between norms:

7.1.3. Definition (Conflicts between norms). Where $n, n' \in Norms$,

Global conflict	Local conflict
$n \perp_{\mathsf{A}} n' := \mathsf{A}\neg(sat(n) \wedge sat(n'))$	$n \perp_{[ac]} n' := [ac]\neg(sat(n) \wedge sat(n'))$

Hence, there is a *global conflict* between two norms when it is logically impossible to jointly fulfill them, and there is a *local conflict* between them when it is impossible, given the circumstances, to jointly fulfill them. So, a norm prescribing that the Doctorate Committee consists of at most six members and a norm prescribing that it consists of at least seven members are globally (hence, locally) in conflict. On the other hand, a norm prescribing that at least half of the voting Committee members are affiliated with the University and a norm prescribing that the voting members shall consist of a majority of full professors are locally (but not globally) in conflict in a situation in which no full professor affiliated with the University is available.

7.1.4. Definition (Implicit prescriptions). Were $n \in Norms$ and $\varphi \in \mathcal{L}_{\mathsf{N}}$,

$$[n]\varphi := \mathsf{A}(sat(n) \rightarrow \varphi)$$

We read $[n]\varphi$ as "norm n implicitly prescribes φ." According to Definition 7.1.4 what is implicitly prescribed by a norm is what is necessary to fulfill it.

In order to capture the basic conception of norms sketched at the beginning of this section, we interpret the language $\mathcal{L}_{\mathsf{N}}$ on frames consisting, at the ontic

level, of a set W of possible states and a relation $R_{[ac]}$ relating every possible state w with the states that are accessible from it (see Chapter 6.2 and 6.3.1). At the deontic level, we assume a function f_{sat} that assigns to every norm n the set of states at which n is fulfilled and a function f_{nor} that assigns to every norm n the set of sentences that make up its explicit content. Importantly, the function f_{nor} does not depend on possible states: this will ensure that the content of a norm remains unchanged across possible states, in line with the assumption that norms are essentially characterized by their content.

7.1.5. Definition (N frame). An N frame is a tuple $\langle W, R_{[ac]}, f_{sat}, f_{nor} \rangle$, where $W \neq \varnothing$, $R_{[ac]} \subseteq W \times W$, $f_{sat} : Norms \to 2^W$, and $f_{nor} : Norms \to 2^{\mathcal{L}_{\mathsf{N}}}$. The functions f_{sat} and f_{nor} are required to satisfy the following conditions: for all $n \in Norms$ and $\varphi \in \mathcal{L}_{\mathsf{N}}$,

1. *Explicit consistency*: if $\varphi \in f_{nor}(n)$, then $\neg\varphi \notin f_{nor}(n)$.
2. *Implicit consistency*: $f_{sat}(n) \neq \varnothing$.
3. *Norm satisfaction*: $sat(n) \in f_{nor}(n)$.
4. *Norm inclusion*: if $sat(n) \in f_{nor}(n')$, then $f_{nor}(n) \subseteq f_{nor}(n')$.

The conditions of N frames reflect our initial assumptions on norms and normative sources. The conditions of explict and implicit consistency ensure that every norm is consistent, both in the explicit sense that no norm explicitly prescribes a proposition and its negation and in the implicit sense that every norm is satisfied at at least one possible state. The condition of norm satisfaction guarantees that every norm prescribes its own satisfaction. Finally, the condition of norm inclusion states that, whenever a norm prescribes the satisfaction of another norm, the explicit content of the former includes the explicit content of the latter.

7.1.6. Definition (N model). An N model is a tuple $\langle F, \nu \rangle$, where F is an N frame and $\nu : Prop \to 2^W$ a valuation function.

7.1.7. Definition (Semantics for $\mathcal{L}_{\mathsf{N}}$). Given an N model M, truth of a formula $\varphi \in \mathcal{L}_{\mathsf{N}}$ at a state w in M, denoted $M, w \models \varphi$, is defined recursively. Truth of atomic propositions and the Boolean connectives is defined as usual. The remaining cases are as follows:

$$
\begin{array}{lll}
M, w \models sat(n) & \text{iff} & w \in f_{sat}(n) \\
M, w \models n : \varphi & \text{iff} & \varphi \in f_{nor}(n) \\
M, w \models \mathsf{A}\varphi & \text{iff} & \text{for all } w' \in W,\ M, w' \models \varphi \\
M, w \models [ac]\varphi & \text{iff} & \text{for all } w' \in W, \text{ if } wR_{[ac]}w', \text{ then } M, w' \models \varphi
\end{array}
$$

In order to provide a sound interpretation to the formulas of $\mathcal{L}_{\mathsf{N}}$, models for N should connect the explicit and implicit content of a norm in a suitable way:

7.1.8. Definition (Suitability of N models). Let $M = \langle W, R_{[ac]}, f_{sat}, f_{nor}, \nu \rangle$ be an N model. Then M is suitable just in case it satisfies the following condition, for all $n \in Norms$, $\varphi \in \mathcal{L}_{\mathsf{N}}$, and $w \in W$:

NS. *Norm Suitability*: if $\varphi \in f_{nor}(n)$ and $w \in f_{sat}(n)$, then $M, w \models \varphi$.

According to Definition 7.1.8, a norm is satisfied only at those states where its explicit prescriptions are satisfied.[2] Together with Definitions 7.1.4 and 7.1.7, this ensures that formulas like $n : \varphi \to [n]\varphi$, which express that the explicit content of a norm is included in its implicit content, are valid in suitable N models.

7.1.2 Axiomatization

The axiom system N is defined by the axioms and rules in Table 7.2. The items in the first two rows and the axioms in group (I) are standard. The axioms in group (II) capture the basic traits of norms: AxN1 codifies the fact that norms are necessarily characterized by their explicit content; AxN2 states that norms are *explicitly* consistent, so that no norm prescribes both φ and $\neg\varphi$; AxN3 reflects the fact that norms are *implicitly* consistent, so that every norm is in principle satisfiable; according to AxN4, norms prescribe at least their own satisfaction; AxN5 has it that the explicit content of a norm prescribing the satisfaction of another norm includes the explicit content of the latter norm; finally, AxN6 corresponds to the condition of norm suitability and expresses that a norm is satisfied only when its explicit prescriptions are fulfilled.

The following theorems will be useful later on.

7.1.9. Theorem. *The following are theorems of* N*:*

1. $n : \varphi \to [n]\varphi$;
2. $n : \varphi \wedge n' : \neg\varphi \to n \perp_{\mathsf{A}} n'$.

Proof:
Theorem 7.1.9(1) follows immediately from AxN1, AxN6, the logic of A and the def. of $[n]\varphi$. Theorem 7.1.9 (2) follows immediately from the same axioms and rules and the def. of $\perp_{\mathsf{A}}$. □

The proof of the following theorem can be derived from the proof of Theorem 7.2.8 below.

7.1.10. Theorem. *The axiom system* N *is sound and strongly complete with respect to class of all suitable* N *models.*

[2] Observe that, since every norm prescribes its own satisfaction [cf. condition 3 in Def. 7.1.5], we also have that, if all $\varphi \in f_{nor}(n)$ are s.t. $M, w \models \varphi$, then $w \in f_{sat}(n)$. Hence, every norm is satisfied *precisely* at those states where its explicitly content is fulfilled.

(CPL)	Classical propositional tautologies	(MP)	From φ and $\varphi \to \psi$, infer ψ
($\mathsf{S5_A}$)	The axiom schemas of S5 for A	($\mathsf{RN_A}$)	From φ, infer $\mathsf{A}\varphi$
	(I) Axioms for $[ac]$		
($\mathsf{K}_{[ac]}$)	$[ac](\varphi \to \psi) \to ([ac]\varphi \to [ac]\psi)$	(Inc)	$\mathsf{A}\varphi \to [ac]\varphi$
	(II) Axioms for $n : \varphi$ **and** $sat(n)$		
(AxN1)	$n : \varphi \to \mathsf{A}(n : \varphi)$	(AxN4)	$n : sat(n)$
(AxN2)	$n : \varphi \to \neg(n : \neg\varphi)$	(AxN5)	$n : sat(n') \wedge n' : \varphi \to n : \varphi$
(AxN3)	$\mathsf{E}(sat(n))$	(AxN6)	$n : \varphi \wedge sat(n) \to \varphi$

Table 7.1: The axiom system N

7.2 The logic of norms and codes NC

In accordance with the intuition that agents use norms to direct their conduct, the aim of the logic of norms and codes NC is to model the obligations generated by *codes of agents*. We take the code of an agent to be the set of norms adopted or accepted by that agent. This can include norms that the agent generally accepts, like the norm prescribing that promises ought to be kept, as well as norms that the agent accepts because of the circumstances, like the norm prescribing to have a travel insurance (accepted when the agent makes vacation plans) or the norm prescribing to pay the monthly rent (accepted when the agent signs a rental agreement). A first difference between agents' codes and norms is thus that the explicit content of an agent's code *can change* across possible states. A second difference is that, since it might consist of norms prescribing incompatible things, an agent's code, unlike a norm, *can be inconsistent*. Finally, we will assume that *agents are cautious* in the sense that they never bind themselves to satisfy the code of another agent, which may change in an unpredictable way. Similarly, we will assume that no norm can prescribe the satisfaction of an agent's code: since the content of a norm is fixed, a norm cannot include prescriptions referring to a normative source whose content might vary.

In this section, we extend the syntax and semantics of the logic N with tools to represent codes of agents (we will refer to them simply as "codes" from now on) and to describe the relations between codes and norms.

7.2.1 Syntax and semantics

Letting $Prop$ and $Norms$ be as before, let us fix a countable set Ag of (names of) agents (we use i, j, k possibly with superscripts $i', i'', \ldots$ for elements of Ag)

and a set *Codes* of (names of) codes. The latter set includes, for every agent i, the code c_i of that agent.

7.2.1. Definition (Syntax of $\mathcal{L}_{\mathsf{NC}}$). Let *Norms*, *Prop*, *Ag*, and *Codes* be defined as above. The set of formulas of $\mathcal{L}_{\mathsf{NC}}$, also denoted $\mathcal{L}_{\mathsf{NC}}$, is generated by the following grammar:

$$\varphi := p \mid sat(n) \mid n : \psi \mid sat(c_i) \mid O_i\chi \mid \neg\varphi \mid (\varphi \wedge \varphi) \mid \mathsf{A}\varphi \mid [ac]\varphi$$

where $p \in Prop$, $n \in Norms$, $c_i \in Codes$, $i \in Ag$, $\psi \in \mathcal{L}_{\mathsf{N}}$, and $\chi \in \mathcal{L}_{\mathsf{N}} \cup \{sat(c_i)\}$. The other logical connectives and the modalities E and $\langle ac \rangle$ are defined as above.

The intended reading of the new formulas is as follows: $sat(c_i)$ means "the code c_i of agent i is satisfied" and $O_i\varphi$ means "φ is obligatory for agent i in virtue of her code." The restrictions we impose on the construction of formulas like $n : \varphi$ and $O_i\varphi$ guarantee that neither the prescriptions of a norm nor the prescriptions of a code contain references to the codes of other agents. On the other hand, notice that formulas like $O_i sat(c_i)$ are allowed, so we can express that a code prescribes its own satisfaction, in accordance with the minimal requirements on normative sources discussed at the beginning of Section 7.1.[3] Besides conflicts between norms [cf. Definition 7.1.3], the extended language allows us to represent conflicts between a norm and a code and between two codes:

7.2.2. Definition (Conflicts involving codes). Where $n \in Norms$ and $c_i, c_j \in Codes$,

Global conflict	Local conflict
$n \perp_{\mathsf{A}} c_i := \mathsf{A}\neg(sat(n) \wedge sat(c_i))$	$n \perp_{[ac]} c_i := [ac]\neg(sat(n) \wedge sat(c_i))$
$c_i \perp_{\mathsf{A}} c_j := \mathsf{A}\neg(sat(c_i) \wedge sat(c_j))$	$c_i \perp_{[ac]} c_j := [ac]\neg(sat(c_i) \wedge sat(c_j))$

More interestingly, we can use formulas like $O_i sat(n)$ to express that *agent i accepts norm n* and formulas like $O_i sat(n) \wedge [ac]O_i sat(n)$ to express that *agent i accepts norm n as binding*. Hence, accepting a norm means to be obliged to satisfy that norm in virtue of one's code. As we will see in a moment, in NC frames this corresponds in a precise sense to including a norm in one's code. Accepting a norm *as binding* means to be obliged to satisfy that norm in virtue of one's code *at all accessible states*. This can be thought of as a form of commitment of the agent to the norm: the agent accepts the norm for now and for the future, rather than just temporarily, i.e., because of the present circumstances. (Contrast, for instance,

[3] Notice that $\mathcal{L}_{\mathsf{NC}}$ does *not* include expressions like $O_iO_i\varphi$ or $O_iO_j\varphi$, even if such expressions seem to have an intuitive reading in natural language (for instance, it makes sense to say that an authority ought to make it obligatory for its subordinates to follow the rules). We will come back to this issue in Section 7.4.2 below.

the acceptance of the norm prescribing to keep promises with the acceptance of the norm prescribing to buy a travel insurance for a two week vacation.)

Turning to the semantics, NC frames differ from N frames in two respects. First, the domain of the function f_{sat} is extended to include the set $Codes$. In this way, f_{sat} will determine, for every normative source, the set of states at which its prescriptions are fulfilled. Second, NC frames include a new function f_{cod} that assigns to every code c_i and possible state w the set of explicit prescriptions of c_i at w. Unlike the function f_{nor}, the function f_{cod} depends on possible states. This allows us to model the fact that the explicit content of a code may change as the circumstances do.

7.2.3. Definition (NC frame). An NC frame is a tuple $\langle W, R_{[ac]}, f_{sat}, f_{nor}, f_{cod}\rangle$, where

- W, $R_{[ac]}$, and f_{nor} are as in Definition 7.1.5;
- $f_{sat} : Norms \cup Codes \to 2^W$;
- $f_{cod} : Codes \times W \to 2^{\mathcal{L}_{\mathsf{NC}}}$ is such that, for all $(c_i, w) \in Codes \times W$, $f_{cod}(c_i, w) \subseteq \mathcal{L}_{\mathsf{N}} \cup \{sat(c_i)\}$.

We require that, besides conditions 1 to 4 from Definition 7.1.5, NC frames satisfy the following conditions: for all $n \in Norms$, $c_i \in Codes$, and $w \in W$,

5. *Code satisfaction*: $sat(c_i) \in f_{cod}(c_i, w)$.
6. *Code inclusion*: if $sat(n) \in f_{cod}(c_i, w)$, then $f_{nor}(n) \subseteq f_{cod}(c_i, w)$.

Thus, like norms, codes prescribe their own satisfaction and, whenever they prescribe the satisfaction of a norm, their explicit content includes all explicit prescriptions of that norm. Note that the above conditions 5 and 6 allow for inconsistencies in the content of a code at some possible states, in line with the idea that an agent might lose track of the accepted norms and end up being guided by conflicting prescriptions.

7.2.4. Definition (NC model). An NC model is a tuple $\langle F, \nu\rangle$, where F is an NC frame and $\nu : Prop \to 2^W$ a valuation function.

7.2.5. Definition (Semantics for $\mathcal{L}_{\mathsf{NC}}$). Given an NC model M, truth of a formula $\varphi \in \mathcal{L}_{\mathsf{NC}}$ at a state w in M, denoted $M, w \models \varphi$, is defined as in Definition 7.1.7 with the addition of the following clauses:

$$\begin{array}{lll} M, w \models sat(c_i) & \text{iff} & w \in f_{sat}(c_i) \\ M, w \models O_i\varphi & \text{iff} & \varphi \in f_{cod}(c_i, w) \end{array}$$

The class of *suitable* NC *models* is defined as follows:

(N)	The axioms and rules of the system N	(AxC2)	$O_i sat(n) \wedge n : \varphi \to O_i\varphi$
(AxC1)	$O_i sat(c_i)$	(AxC3)	$O_i\varphi \wedge sat(c_i) \to \varphi$

Table 7.2: The axiom system NC

7.2.6. Definition (Suitability of NC models). An NC model

$$M = \langle W, R_{[ac]}, f_{sat}, f_{nor}, f_{cod}, \nu \rangle$$

is suitable just in case it satisfies condition NS from Definition 7.1.8 plus the following condition, for all $c_i \in Codes$, $\varphi \in \mathcal{L}_{\mathsf{N}}$, and $w \in W$:

CS. *Code Suitability*: if $\varphi \in f_{cod}(c_i, w)$ and $w \in f_{sat}(c_i)$, then $M, w \models \varphi$.

7.2.2 Axiomatization

The axiom system NC is defined as shown in Table 7.2. Axioms AxC1 to AxC3 parallel axioms AxN4 to AxN6 from Table 7.1 and characterize codes as normative sources satisfying the minimal requirements discussed at the beginning of Section 7.1. Observe that AxC2 can be read as saying that whenever an agent accepts a norm, all prescriptions of that norm become obligatory for him in virtue of his code. Hence, the code of an agent includes all the norms she accepts by including their explicit content. The following theorems will be important later on.

7.2.7. Theorem. *The following are theorems of* NC*:*

1. $[ac]O_i\varphi \wedge [ac]O_j\neg\varphi \to c_i \perp_{[ac]} c_j$;
2. $[ac]O_i sat(n) \wedge [ac]O_j sat(n') \wedge n \perp_{[ac]} n' \to c_i \perp_{[ac]} c_j$.

Proof:
Both theorems follow immediately from AxC3, the logic of $[ac]$, and the def. of $\perp_{[ac]}$. □

7.2.8. Theorem. *The axiom system* NC *is sound and strongly complete with respect to class of all suitable* NC *models.*

The proof of soundness is a matter of routine validity check. The proof of completeness is based on the construction of a canonical model for NC and, by paying attention to the universal modality, proceeds in an entirely standard way [see Blackburn et al., 2001, Chapter 4.2; Goranko and Passy, 1992]. We only provide the definition of the canonical model for NC and leave the details of the proof to the reader. As usual, let $\mathcal{W}$ be the set of all maximal consistent sets of NC. Where $w \in \mathcal{W}$ define: $w/[ac] = \{\varphi \in \mathcal{L}_{\mathsf{NC}} \mid [ac]\varphi \in w\}$.

7.2.9. Definition (Canonical $\mathsf{PD}_e\mathsf{LO}$ model for w_0). The canonical NC model for $w_0 \in \mathcal{W}$ is a tuple $M^c = \left\langle W^c, R^c_{[ac]}, f^c_{sat}, f^c_{nor}, f^c_{cod}, \nu^c \right\rangle$, where

- $W^c = \{w \in \mathcal{W} \mid w_0/\mathsf{A} \subseteq w\}$;
- $R^c_{[ac]} \subseteq W^c \times W^c$ is such that, for all $w, w' \in W^c$, $wR^c_{[ac]}w'$ iff $w/[ac] \subseteq w'$;
- $f^c_{sat} : Norms \cup Codes \to 2^{W^c}$ is such that, for all $x \in Norms \cup Codes$, $w \in f^c_{sat}(x)$ iff $sat(x) \in w$;
- $f^c_{nor} : Norms \to 2^{\mathcal{L}_\mathsf{N}}$ is such that, for all $n \in Norms$ and $\varphi \in \mathcal{L}_\mathsf{N}$, $\varphi \in f^c_{nor}(n)$ iff $n : \varphi \in w_0$;
- $f^c_{cod} : Codes \times W^c \to 2^{\mathcal{L}_\mathsf{NC}}$ iff, for all $(c_i, w) \in Codes \times W^c$ and $\varphi \in \mathcal{L}_\mathsf{NC}$, $\varphi \in f^c_{cod}(c_i, w)$ iff $O_i\varphi \in w$;
- $\nu^c : Prop \to 2^{W^c}$ is such that, for all $w \in W^c$, $w \in \nu^c(p)$ iff $p \in w$.

Given Definition 7.2.9, the proof of the Truth Lemma (for every $w \in W^c$ and $\varphi \in \mathcal{L}_\mathsf{NC}$, $M^c, w \models \varphi$ iff $\varphi \in w$) and the proof that M^c is a NC model proceed in the usual way.[4] This is sufficient to conclude that every consistent set $\Gamma \subseteq \mathcal{L}_\mathsf{NC}$ is satisfiable in an NC model.

7.3 Updating codes: the dynamic system DNC

The logic NC is suitable to model the basic relations between norms and codes, but it is not suitable to model the fact that, when an agent finds herself in a new situation, she typically needs to specify her code by adding new norms. In fact, a code is often not specific enough to determine what the agent ought to do with respect to all decisions she has to take. Here is a simple example:

7.3.1. Example. While going to the office, Carl assists to a car accident. Carl's code does not determine whether, under the unexpected circumstances, he should give first aid or ignore the fact and go straight to the office. If his code is to direct his behavior, Carl has to accept a new norm.

A natural way to improve our framework is to exploit the idea underlying DEL [Baltag et al., 1998; van Benthem, 2011; van Ditmarsch et al., 2008]: we can view the specification of a code as an update procedure that takes a model representing the initial deontic situation of the agent and returns the model representing the updated deontic situation in which the agent has included a new norm in her code. Let us start by extending $\mathcal{L}_\mathsf{NC}$ with dynamic modalities for code update $[c_i \oplus n]$ allowing us to describe what is true after agent i has updated her code with norm n.

[4] Let us only mention that the definition of f^c_{cod} ensures that $f^c_{cod}(c_i, w) \subseteq \mathcal{L}_\mathsf{N} \cup \{sat(c_i)\}$ because of the restrictions we imposed on $\mathcal{L}_\mathsf{NC}$: $O_i\varphi \in \mathcal{L}_\mathsf{NC}$ just in case $\varphi \in \mathcal{L}_\mathsf{N} \cup \{sat(c_i)\}$.

7.3.2. Definition (Syntax of $\mathcal{L}_{\mathsf{DNC}}$). Let $Norms$, Ag, $Codes$ and $Prop$ be as in Definitions 7.1.2 and 7.2.1. The set of formulas of DNC, also denoted $\mathcal{L}_{\mathsf{DNC}}$, is generated by the following grammar:

$$\varphi := p \,|\, sat(n) \,|\, n : \psi \,|\, sat(c_i) \,|\, O_i\chi \,|\, \neg\varphi \,|\, (\varphi \wedge \varphi) \,|\, \mathsf{A}\varphi \,|\, [ac]\varphi \,|\, [c_i \oplus n]\varphi$$

where, $p \in Prop$, $n \in Norms$, $i \in Ag$, $c_i \in Codes$, and ψ and χ are as in Def. 7.2.1 with the additional constraint that no dynamic modality occurs in them.

Formulas like $[c_i \oplus n]\varphi$ mean "after agent i has accepted norm n, φ is true."

The semantics for the dynamic modalities requires the definition of an update procedure modeling the changes deriving from the inclusion of a norm in a code. There are two key changes: first, at every state, the *explicit content* of the code in question is extended so as to include all the explicit prescriptions of the accepted norm; second, the set of states at which the code in question is *satisfied* is restricted to the set of states at which both the original code and the accepted norm are satisfied. The following definition encodes the described procedure of code update.

7.3.3. Definition (Update model $M^{c_i\oplus n}$). Let $c_i \in Codes$, $n \in Norms$, and $M = \langle W, R_{[ac]}, f_{sat}, f_{nor}, f_{cod}, \nu\rangle$ be an NC model. The update of M by c_i and n is the tuple $M^{c_i\oplus n} = \left\langle W^{c_i\oplus n}, R^{c_i\oplus n}_{[ac]}, f^{c_i\oplus n}_{sat}, f^{c_i\oplus n}_{nor}, f^{c_i\oplus n}_{cod}, \nu^{c_i\oplus n}\right\rangle$ where:

- $W^{c_i\oplus n} = W$, $R^{c_i\oplus n}_{[ac]} = R_{[ac]}$, and $\nu^{c_i\oplus n} = \nu$;
- $f^{c_i\oplus n}_{sat}(x) = \begin{cases} f^{c_i\oplus n}_{sat}(x) & \text{if } x \neq c_i \\ f_{sat}(x) \cap f_{sat}(n) & \text{if } x = c_i \end{cases}$
- $f^{c_i\oplus n}_{cod}(x,w) = \begin{cases} f^{c_i\oplus n}_{cod}(x,w) & \text{if } x \neq c_i \\ f_{cod}(x,w) \cup f_{nor}(n) & \text{if } x = c_i \end{cases}$

7.3.4. Proposition. *Let $c_i \in Codes$ and $n \in Norms$. For any* NC *model M, the update model $M^{c_i\oplus n}$ is an* NC *model.*

Proof:
Since the procedure of code update only affects the functions f_{sat} and f_{cod} when applied to c_i and there is no condition on the restriction of f_{sat} to $Codes$ characterizing NC models, we only need to check that $f^{c_i\oplus n}_{cod}$ satisfies the requirements in Def. 7.2.3 when applied to c_i. Since M is an NC model, we have that: (1) $f_{cod}(c_i, w) \subseteq \mathcal{L}_{\mathsf{N}} \cup \{sat(c_i)\}$ and $f_{nor}(n) \subseteq \mathcal{L}_{\mathsf{N}}$. Hence, $f^{c_i\oplus n}_{cod}(c_i, w) = f_{cod}(c_i, w) \cup f_{nor}(n) \subseteq \mathcal{L}_{\mathsf{N}} \cup \{sat(c_i)\}$, i.e., $f^{c_i\oplus n}_{cod}$ is well-defined. (2) $sat(c_i) \in f_{cod}(c_i, w)$ by the condition of code satisfaction. Hence, $sat(c_i) \in f_{cod}(c_i, w) \cup f_{nor}(n) = f^{c_i\oplus n}_{cod}(c_i, w)$, i.e., $M^{c_i\oplus n}$ satisfies the condition of code satisfaction. (3) Take any norm n' s.t. $sat(n') \in f^{c_i\oplus n}_{cod}(c_i, w) = f_{cod}(c_i, w) \cup f_{nor}(n)$. If

$sat(n') \in f_{cod}(c_i, w)$, then $f_{nor}(n') \subseteq f_{cod}(c_i, w)$ by the condition of code inclusion. If $sat(n') \in f_{nor}(n)$, then $f_{nor}(n') \subseteq f_{nor}(n)$ by the condition of norm inclusion. Either case, $f_{nor}(n') \subseteq f_{cod}^{c_i \oplus n}(c_i)$, i.e., $M^{c_i \oplus n}$ satisfies the condition of code inclusion. □

The semantics for the dynamic modalities is now defined in the usual way:

7.3.5. Definition (Semantics for $\mathcal{L}_{\mathsf{DNC}}$). Given an NC model M, truth of a formula $\varphi \in \mathcal{L}_{\mathsf{DNC}}$ at a state w in M, denoted $M, w \models \varphi$, is defined as in Def. 7.2.5 with the addition of the following new case:

$$M, w \models [c_i \oplus n_i]\varphi \quad \text{iff} \quad M^{c_i \oplus n}, w \models \varphi$$

The following lemma states that the satisfaction of formulas in $\mathcal{L}_{\mathsf{N}}$ is invariant under the procedure of code update. So, unsurprisingly, formulas in $\mathcal{L}_{\mathsf{N}}$ cannot distinguish between a model and an update model obtained from it.

7.3.6. Lemma. *Let M be an* NC *model. Then, for any $\varphi \in \mathcal{L}_{\mathsf{N}}$ and $w \in W$, $M, w \models \varphi$ iff $M^{c_i \oplus n} \models \varphi$.*

Proof:
The proof is by induction on the complexity of $\varphi \in \mathcal{L}_{\mathsf{N}}$ and follows straightforwardly from the fact that the procedure of code update described in Def. 7.3.3 does not affect the semantic components needed to evaluate formulas in $\mathcal{L}_{\mathsf{N}}$. □

7.3.7. Proposition. *Let $c_i \in Codes$ and $n \in Norms$. For any suitable* NC *model M, the update model $M^{c_i \oplus n}$ is a suitable* NC *model.*

Proof:
By Prop. 7.3.4, we already know that $M^{c_i \oplus n}$ is an NC model, so we only need to show that $M^{c_i \oplus n}$ satisfies the suitability conditions stated in Def. 7.1.8 and Def. 7.2.6:

NS If $\varphi \in f_{nor}^{c_i \oplus n}(n')$ and $w \in f_{sat}^{c_i \oplus n}(n')$, then $M^{c_i \oplus n}, w \models \varphi$.

Assume the antecedent. Since $f_{nor}^{c_i \oplus n}(n') = f_{nor}(n')$ and $f_{sat}^{c_i \oplus n}(n') = f_{sat}(n')$ by Def. 7.3.3, $M, w \models \varphi$, as M is suitable by hypothesis. Hence, $M^{c_i \oplus n}, w \models \varphi$ by Lem. 7.3.6 (since $\varphi \in f_{nor}(n') \subseteq \mathcal{L}_{\mathsf{N}}$).

CS If $\varphi \in f_{cod}^{c_i \oplus n}(c_j, w)$ and $w \in f_{sat}^{c_i \oplus n}(c_j)$, then $M^{c_i \oplus n}, w \models \varphi$.

The relevant case is when $c_j = c_i$. Assume the antecedent. By Def. 7.3.3, $\varphi \in f_{cod}(c_i, w) \cup f_{nor}(n)$ and $w \in f_{sat}(c_i) \cap f_{sat}(n)$. There are two cases. (I) If $\varphi \in f_{cod}(c_i, w)$, then (1) $M, w \models \varphi$ since $w \in f_{sat}(c_i)$ and M is suitable by hypothesis and (2) $\varphi \in \mathcal{L}_{\mathsf{N}} \cup \{sat(c_i)\}$. If $\varphi \in \mathcal{L}_{\mathsf{N}}$, then (1) implies that $M^{c_i \oplus n}, w \models \varphi$ by Lem. 7.3.6. If φ is $sat(c_i)$, then $M^{c_i \oplus n}, w \models \varphi$ by Def. 7.3.5, as $w \in f_{sat}^{c_i \oplus n}(c_i)$ by hypothesis. (II) If $\varphi \in f_{nor}(n)$, then (1) $M, w \models \varphi$ since $w \in f_{sat}(n)$ and M is suitable by hypothesis and (2) $\varphi \in \mathcal{L}_{\mathsf{N}}$. Hence, $M^{c_i \oplus n}, w \models \varphi$ by Lem. 7.3.6.

(I) Ontic formulas

R_p	$[c_i \oplus n]p \leftrightarrow p$	R_{A}	$[c_i \oplus n]\mathsf{A}\varphi \leftrightarrow \mathsf{A}[c_i \oplus n]\varphi$
$\mathsf{R}_\neg$	$[c_i \oplus n]\neg\varphi \leftrightarrow \neg[c_i \oplus n]\varphi$	$\mathsf{R}_{[ac]}$	$[c_i \oplus n][ac]\varphi \leftrightarrow [ac][c_i \oplus n]\varphi$
$\mathsf{R}_\wedge$	$[c_i \oplus n](\varphi \wedge \psi) \leftrightarrow$ $[c_i \oplus n]\varphi \wedge [c_i \oplus n]\psi$	$\mathsf{RN}_{[c_i \oplus n]}$	from φ, infer $[c_i \oplus n]\varphi$

(II) Deontic formulas

R_n	$[c_i \oplus n]n' : \varphi \leftrightarrow n' : \varphi$	$\mathsf{R}_{sat(n)}$	$[c_i \oplus n]sat(n') \leftrightarrow sat(n')$
R_{O_j}	$[c_i \oplus n]O_j\varphi \leftrightarrow O_j\varphi$, for $j \neq i$	$\mathsf{R}_{sat(c_j)}$	$[c_i \oplus n]sat(c_j) \leftrightarrow sat(c_j)$, for $j \neq i$
R_{O_i}	$[c_i \oplus n]O_i\varphi \leftrightarrow O_i\varphi \vee n : \varphi$	$\mathsf{R}_{sat(c_i)}$	$[c_i \oplus n]sat(c_i) \leftrightarrow sat(c_i) \wedge sat(n)$

Table 7.3: Reduction axioms and inference rule for $[c_i \oplus n]$

□

Propositions 7.3.4 and 7.3.7 tell us that the procedure of code update does not bring us outside the class of suitable NC models. We can now prove the central theorem of this section.

7.3.8. Theorem. *The axiom system* DNC*, defined by extending the axiom system* NC *with the axioms and rules in Table 7.3, is sound and complete with respect to the class of all suitable* NC *models.*

Proof:
The proof of soundness follows immediately from soundness of NC [Thm. 7.2.8] and the fact that, as the reader can easily verify, the axioms and rules in Table 7.3 are, resp., valid and truth preserving in the class of NC models. The proof of completeness proceeds by using standard techniques from DEL [see, e.g., van Benthem et al., 2006]: Given the reduction axioms in Table 7.3, it is not difficult to prove that every formula in $\mathcal{L}_{\mathsf{DNC}}$ is provably equivalent to a formula in $\mathcal{L}_{\mathsf{NC}}$. So, if $\varphi \in \mathcal{L}_{\mathsf{DNC}}$ is valid in the class of suitable NC models, then a provably equivalent $\varphi' \in \mathcal{L}_{\mathsf{NC}}$ is valid in it. Since NC is complete w.r.t. the class of suitable NC models [cf. Thm. 7.2.8], it follows that φ' is provable in NC. Since NC is included in DNC, we conclude that φ' (hence, φ) is provable in DNC. □

While the axioms in group (I) are standard, the axioms in group (II) reflect the fact that the procedure of code update affects the deontic components of an NC model only when they are applied to the updated code. We will use the following theorem in the next section.

7.3.9. THEOREM. *The following are theorems of* DNC*:*

1. $[c_i \oplus n]O_i sat(n)$
2. $n : \varphi \rightarrow [c_i \oplus n]\mathsf{A}O_i\varphi$
3. $n : \varphi \rightarrow [c_i \oplus n][ac]O_i\varphi$

Proof:
Thm. 7.3.9 (1) follows immediately from AxN4 and $\mathsf{RN}_{c_i \oplus n}$. Thm. 7.3.9 (2) is an immediate consequence of AxN1, R_A and R_{O_i}. Similarly, Thm. 7.3.9 (3) follows from AxN1, R_{ac} and R_{O_i}. □

Hence, after agent i specifies her code by including norm n, she accepts n and everything that is prescribed by n, as it should be.

7.4 Applications and an extension

In this section, we illustrate how the system DNC can be used to provide an insightful analysis of, and to draw important distinctions between, normative conflicts involving more than one agent. We start by modeling two well-known paradigmatic examples, namely those of Antigone and Gandhi. We then show how DNC and a simple variant of it allow us to capture key traits distinguishing cases of conscientious objection from cases of civil disobedience.

7.4.1 Keeping track of the source of a conflict

7.4.1. EXAMPLE. In Sophocles's tragedy, Antigone accepts the law of the gods, according to which her brother Polynices ought to be buried. Yet, according to the law of the state, enacted by Creon, Polynices ought not to be buried.

Example 7.4.1 presents two normatively relevant agents, namely Antigone (let c_1 be her code) and Creon, who is associated with the code of the state (let it be c_2). The norms of interest are the norm of the gods (n) prescribing that Polynices is buried ($n : \varphi$) and the conflicting norm n' promulgated by Creon and prescribing that Polynices is left unburied ($n' : \neg\varphi$). Since the prescriptions of the two norms are jointly inconsistent, there is a global conflict between them ($n \perp_\mathsf{A} n'$). We can use the dynamics of DNC to see how this conflict generates a local conflict between the codes c_1 and c_2. In line with the story, let us assume that Antigone accepts the norm of the gods n as binding ($O_1 sat(n) \wedge [ac]O_1 sat(n)$). DNC can then represent the fact that, by updating the code of the state with the conflicting norm n', Creon generates a conflict between the codes c_1 and c_2. More specifically, letting $\Gamma = \{n : \varphi, n' : \neg\varphi, O_1 sat(n), [ac]O_1 sat(n)\}$, we can prove that $[c_2 \oplus n']c_2 \perp_{[ac]} c_1$ is deducible from Γ in DNC. The proof is as follows:

(1)	$\Gamma \vdash_{\mathsf{DNC}} n : \varphi \wedge n' : \neg\varphi$	by logic
(2)	$\Gamma \vdash_{\mathsf{DNC}} [ac]O_1 sat(n)$	by logic
(3)	$\Gamma \vdash_{\mathsf{DNC}} [ac]O_1\varphi$	from (1), (2) and AxC2
(4)	$\Gamma \vdash_{\mathsf{DNC}} [c_2 \oplus n'][ac]O_1\varphi$	from (3) and $\mathsf{R}_{[ac]}$
(5)	$\Gamma \vdash_{\mathsf{DNC}} [c_2 \oplus n'][ac]O_2\neg\varphi$	from (1) and Thm. 7.3.9(3)
(6)	$\Gamma \vdash_{\mathsf{DNC}} [c_2 \oplus n']c_2 \perp_{[ac]} c_1$	from (4) and (5), Thm. 7.2.7(1) and the logic of $[c_2 \oplus n']$

What is particularly interesting is that the dynamic operator $[c_2 \oplus n']$ can be used to represent the origin of the conflict by keeping track of the fact that the clash between the two codes is due to Creon's decision to change the legal code rather than to Antigone's choice. Gandhi's case differs in this respect.

7.4.2. Example. Gandhi explicitly opposed the colonial rules imposed by the British Empire by employing a non-violent form of civil disobedience.

As Example 7.4.1, Example 7.4.2 presents two normatively relevant agents, namely Gandhi (let c_3 be his code) and the British Empire (let c_4 be its code). We can think of the colonial rules opposed by Gandhi as the prescriptions of a norm n establishing and regulating the colonial status of India. While the legal code of the British Empire includes the norm n as binding $(O_4 sat(n) \wedge [ac]O_4 sat(n))$, Gandhi's code commits him to violate n $(O_3\neg sat(n) \wedge [ac]O_3\neg sat(n))$. It is immediately seen that this gives rise to a local conflict between the two codes c_3 and c_4.

The example becomes more interesting when analyzed from a dynamic perspective. Suppose that Gandhi started opposing the colonial rules when he realized what it meant for India to be a British colony. We can then represent the origin of the conflict by introducing a norm n' prescribing to violate the colonial rules encoded by the norm n $(n' : \neg sat(n))$. We can use the resources of DNC to represent the fact that the conflict between Gandhi and the British Empire arises when Gandhi specifies his code by accepting n'. Letting $\Sigma = \{n' : \neg sat(n), O_4 sat(n), [ac]O_4 sat(n)\}$, we can prove that $[c_3 \oplus n']c_3 \perp_{[ac]} c_4$ is deducible from Σ in DNC. The proof is as follows:

(1)	$\Sigma \vdash_{\mathsf{DNC}} n : sat(n) \wedge n' : \neg sat(n)$	by logic and AxN4
(2)	$\Sigma \vdash_{\mathsf{DNC}} n \perp_{\mathsf{A}} n'$	from (1) by Thm. 7.2.7(2)
(3)	$\Sigma \vdash_{\mathsf{DNC}} n \perp_{[ac]} n'$	from (2) and Inc
(4)	$\Sigma \vdash_{\mathsf{DNC}} [c_3 \oplus n']n \perp_{[ac]} n'$	from (3), $\mathsf{R}_\neg$, $\mathsf{R}_{[ac]}$, R_{sat}
(5)	$\Sigma \vdash_{\mathsf{DNC}} [ac]O_4 sat(n)$	by logic
(6)	$\Sigma \vdash_{\mathsf{DNC}} [c_3 \oplus n'][ac]O_4 sat(n)$	from (5) and $\mathsf{R}_{[ac]}$
(7)	$\Sigma \vdash_{\mathsf{DNC}} [c_3 \oplus n'][ac]O_3\neg sat(n)$	from (1) and Thm. 7.3.9(3)
(8)	$\Sigma \vdash_{\mathsf{DNC}} [c_3 \oplus n']c_3 \perp_{[ac]} c_4$	from (4), (6), (7), Thm. 7.2.7(1) and the logic of $[c_2 \oplus n']$

In this case the conflict can thus be represented as depending on Gandhi's decision to violate the laws of the state rather than on a change in the latter laws. This suggests that DNC is suitable to model a basic feature of civil disobedience as opposed to conscientious objection: A civil disobedient, like Gandhi, overtly opposes the current laws. On the other hand, a conscientious objector, like Antigone, at first opposes the current laws because the latter turn out to be wrong given her code, and not because she explicitly accepts a norm prescribing to oppose these laws. In other words, DNC seems to have the resources to account for the fact that the origin of the opposition to the state are different in the two cases. Let us explore a bit further how this distinction can be analyzed for in DNC.

7.4.2 Civil disobedience and conscientious objection

Despite being complex phenomena, cases of *civil disobedience* essentially involve three key elements [see, e.g., Brownlee, 2012; Smith, 2013]. That is:

(C) *Conscientiousness*: a civil disobedient thinks that the laws of the state clash with the right conception of good or justice;

(F) *Faithfulness to the law*: a civil disobedient is willing to accept the right laws and to communicate with the government;

(A) *Constructive aim*: a civil disobedient aims at changing the laws of the state rather than overturning the entire legal system.

Although they also involve conscientiousness, cases of *conscientious objection* might fail to involve faithfulness to the law or serve a constructive aim.

The analysis of these elements in DNC rests on the assumption that a code consists of the norms that better capture the agent's conception of justice or, more generally, what the agent takes to be "deontically ideal." Under this assumption, we can read $O_i\varphi$ as saying that φ is obligatory for i in virtue of her deontic ideal. We can then represent at least the first two components of civil disobedience using formulas like the following:

(A) $O_{state}\varphi \wedge O_i\psi \wedge \varphi \perp_{[ac]} \psi$;

(F) $O_{state}\varphi \wedge \langle ac \rangle (\varphi \wedge sat(c_i)) \rightarrow O_i\varphi$.

(A) says that the ideal acknowledged by the state prescribes something that clashes with what the ideal acknowledged by agent i prescribes. On the other hand, (F) says that the ideal acknowledged by agent i prescribes that i obeys all the prescriptions of the state that do not clash with the ideal itself.[5]

[5]Our analysis of faithfulness to the law thus represents faithfulness *pro tanto*: civil disobedients are willing to respect the current laws only to the extent that they do not clash with their conception of good or justice.

What about the last component of civil disobedience, i.e., serving a constructive aim? Suppose that agent i is in a situation satisfying (A) for some φ and ψ. Then, there are two natural ways to express that i aims at changing the laws of the state, namely:

$(A.1)$ $O_i\neg O_{state}\varphi$ (negative aim);

$(A.2)$ $O_iO_{state}\psi$ (positive aim).

While $(A.1)$ says that, given her deontic ideal, i ought to make the code of the state such that φ ceases to be obligatory for the state, $(A.2)$ says that, given her deontic ideal, i ought to make the code of the state such that ψ becomes obligatory for the state. If these formulas were available in $\mathcal{L}_{\mathsf{DNC}}$, then "serving a constructive aim" could be broken down into "serving a negative aim" and "serving a positive aim." Letting $P_i\varphi$ abbreviate $\neg O_i\neg\varphi$, we could also express that an agent is serving what we might call a *neutral aim* by means of the formula $O_iP_{state}\psi$. This would say that i ought to change c_{state} in such a way that the state at least ceases to prohibit ψ.

Now, nesting of obligation operators is not allowed in $\mathcal{L}_{\mathsf{DNC}}$, so the suggested representations of constructive (negative, positive, or neutral) aims are not available in DNC. Yet, it turns out that our dynamic system can be refined so as to allow the amount of nesting of obligation operators needed to model constructive aims. We devote the rest of this section to the presentation of this refinement. We start by introducing the static system NC^+.

The system NC^+

The the set of formulas of the language $\mathcal{L}_{\mathsf{NC}^+}$ is defined as the set of formulas of $\mathcal{L}_{\mathsf{NC}}$ except that more formulas are allowed in the scope of the deontic operators O_i. Specifically, let:

$$\mathcal{O} = \{O_i\varphi \mid i \in Ag \text{ and } \varphi \in \mathcal{L}_{\mathsf{N}}\} \qquad \overline{\mathcal{O}} = \{\neg O_i\varphi \mid i \in Ag \text{ and } \varphi \in \mathcal{L}_{\mathsf{N}}\}$$

We require that in a formula like $O_i\varphi$, $\varphi \in \mathcal{L}_{\mathsf{N}} \cup \{sat(c_i)\} \cup \mathcal{O} \cup \overline{\mathcal{O}}$. Accordingly, formulas of form $O_iO_j\varphi$ and $O_i\neg O_j\varphi$ are now allowed, even for $i \neq j$. Intuitively, $O_iO_j\varphi$ says that agent i ought to make the code of agent j such that φ becomes obligatory for j, while $O_i\neg O_j\varphi$ says that i ought to make the code of j such that φ ceases to be obligatory for j. In order to interpret formulas of $\mathcal{L}_{\mathsf{NC}^+}$, we adapt the notion of NC frame by relaxing the conditions on f_{cod}:

7.4.3. Definition (NC^+ frame). An NC^+ frame is a tuple

$$\langle W, R_{[ac]}, f_{sat}, f_{nor}, f_{cod}\rangle$$

where W, $R_{[ac]}$, f_{sat}, f_{nor}, ν are as in Definition 7.2.3 and $f_{cod} : Codes \times W \to 2^{\mathcal{L}_{\mathsf{NC}}}$ is such that, for all $(c_i, w) \in Codes \times W$, $f_{cod}(c_i, w) \subseteq \mathcal{L}_{\mathsf{N}} \cup \{sat(c_i)\} \cup \mathcal{O} \cup \overline{\mathcal{O}}$.

NC^+ frames satisfy all conditions stated in Definition 7.2.3 plus the following condition: for all $n \in Norms$, $j \in Ag$, $(c_i, w) \in Codes \times W$, and $\varphi \in \mathcal{L}_{\mathsf{N}}$,

6. *Coherence of Codes*: if $\neg O_j\varphi \in f_{cod}(c_i, w)$ and $\varphi \in f_{nor}(n)$, then $\neg O_j sat(n) \in f_{cod}(c_i, w)$.

The condition of coherence of codes says that, if a code c_i requires that an agent j is not obliged to realize φ in virtue of her code, then c_i also requires that j does not accept any norm prescribing φ. An NC^+ model is an NC^+ frame supplemented with a valuation function. The notions of truth and suitable model for $\mathcal{L}_{\mathsf{NC}^+}$ are defined as in Definitions 7.2.5 and 7.2.6. It is not difficult to see that the system NC^+ obtained by extending NC with the axiom schema

(AxC4) $O_i\neg O_j\varphi \wedge n : \varphi \rightarrow O_i\neg O_j sat(n)$

is sound and complete with respect to the class of all suitable NC^+ models.

Dynamics

We now extend the language $\mathcal{L}_{\mathsf{NC}^+}$ with dynamic modalities $[c_i \boxplus n]$ as we did in Section 7.3. In order to interpret the extended language $\mathcal{L}_{\mathsf{DNC}^+}$ we modify Definition 7.3.3 in order to accommodate the fact that, by accepting a new norm n, an agent i affects the satisfaction of all codes that prescribe that i ought not to accept n.

7.4.4. Definition (Modified update model $M^{c_i \boxplus n}$). Let $c_i \in Codes$, $n \in Norms$, and $M = \langle W, R_{[ac]}, f_{sat}, f_{nor}, f_{cod}, \nu\rangle$ be an NC^+ model. The modified update of M by c_i and n is the tuple $M^{c_i \boxplus n} = \left\langle W^{c_i \boxplus n}, R^{c_i \boxplus n}_{[1]}, f^{c_i \boxplus n}_{sat}, f^{c_i \boxplus n}_{nor}, f^{c_i \boxplus n}_{cod}, \nu^{c_i \boxplus n}\right\rangle$, where all elements are defined as in Definition 7.3.3, except for $f^{c_i \boxplus n}_{sat}$ which is defined as follows:

$$f^{c_i \boxplus n}_{sat}(x) = \begin{cases} f^{c_i \oplus n}_{sat}(x) & \text{if } x \in Norms \\ f_{sat}(x) \cap \{w \in W \mid \neg O_i sat(n) \notin f_{cod}(x, w)\} & \text{if } x \in Codes \setminus \{c_i\} \\ f_{sat}(x) \cap f_{sat}(n) \cap \{w \in W \mid \neg O_i sat(n) \notin f_{cod}(x, w)\} & \text{if } x = c_i \end{cases}$$

The evaluation rule for $[c_i \boxplus n]\varphi$ is then as expected:

$$M, w \models [c_i \boxplus n]\varphi \text{ iff } M^{c_i \boxplus n}, w \models \varphi$$

where M is an NC^+ model and w a state in it. As before, the following lemma is an immediate consequence of the fact that the update procedure described in Definition 7.4.4 does not affect the semantic components needed to evaluate formulas in $\mathcal{L}_{\mathsf{N}}$.

7.4.5. Lemma. *Let M be an NC^+ model. Then, for any $\varphi \in \mathcal{L}_{\mathsf{N}}$ and $w \in W$, $M, w \models \varphi$ iff $M^{c_i \boxplus n} \models \varphi$.*

7.4.6. PROPOSITION. *Let $n \in Norms$ and $c_i \in Codes$. For any suitable NC^+ model M, the modified update model $M^{c_i \boxplus n}$ is a suitable NC^+ model.*

Proof:
(I) The proof that $M^{c_i \boxplus n}$ is an NC^+ model proceeds as the proof of Prop. 7.3.4, except that now we need to show that the condition of coherence of codes is satisfied. So, suppose that (1) $\neg O_j \varphi \in f_{cod}^{c_i \boxplus n}(c_i, w)$ and (2) $\varphi \in f_{nor}^{c_i \boxplus n}(n)$. We need to show that $\neg O_j sat(n) \in f_{cod}^{c_i \boxplus n}(c_i, w)$. By (1) and Def. 7.4.4, $\neg O_j \varphi \in f_{cod}(c_i, w) \cup f_{nor}(n)$. Since $\neg O_j \varphi \notin \mathcal{L}_{\mathsf{N}}$ and $f_{nor}(n) \subseteq \mathcal{L}_{\mathsf{N}}$, (3) $\neg O_j \varphi \in f_{cod}(c_i, w)$. By the condition of coherence of codes, (3) and (2) imply $\neg O_j sat(n) \in f_{cod}(c_i, w)$. Hence, $\neg O_j sat(n) \in f_{cod}^{c_i \boxplus n}(c_i, w)$ by the def. of $f_{cod}^{c_i \boxplus n}$. (II) Given Lem. 7.4.5 the proof that $M^{c_i \boxplus n}$ satisfies the condition NS of norm suitability proceeds as in the proof of Thm. 7.3.7. For the condition CS of code suitability, we need to consider the case in which $c_i \neq c_j$ and the case in which $c_i = c_j$. We only present the proof for the latter case (the proof for the former proceeds in a similar way). Suppose that (1) $\varphi \in f_{cod}^{c_i \boxplus n}(c_i, w)$ and (2) $w \in f_{sat}^{c_i \boxplus n}(c_i)$. We need to show that $M^{c_i \boxplus n}, w \models \varphi$. By Def.7.4.4, (1') $\varphi \in f_{cod}(c_i, w) \cup f_{nor}(n)$ and (2') $w \in f_{sat}(c_i) \cap f_{sat}(n) \cap \{w \in W \mid \neg O_i sat(n) \notin f_{cod}(c_i, w)\}$. In turn, by the def. of $f_{cod}(c_i, w)$ and $f_{nor}(n)$, (1") $\varphi \in \mathcal{L}_{\mathsf{N}} \cup \{sat(c_i)\} \cup \mathcal{O} \cup \overline{\mathcal{O}}$. The interesting case is when $\varphi \in \overline{\mathcal{O}}$. So, let us assume that φ has form $\neg O_j \psi$, for some $j \in Ag$ and $\psi \in \mathcal{L}_{\mathsf{N}}$. Since $\neg O_j \psi \in f_{cod}(c_i, w)$ by (1) and $w \in f_{sat}(c_i)$ by (2'), $M, w \models \neg O_j \psi$, as M is suitable by hypothesis. By the def. of truth, this means that (3) $\psi \notin f_{cod}(c_j)$. Now, if $j \neq i$, then $f_{cod}(c_j) = f_{cod}^{c_i \boxplus n}(c_j)$, and so $M^{c_i \boxplus n}, w \models \neg O_j \psi$, as desired. If $j = i$, then we also need to show that $\psi \notin f_{nor}(n)$. Suppose, toward contradiction, that (4) $\psi \in f_{nor}(n)$. Since we assumed that $\neg O_j \psi \in f_{cod}(c_i, w)$, (4) implies that $\neg O_j sat(n) \in f_{cod}(c_i, w)$ by the condition of coherence of codes. But, by (2'), $\neg O_j sat(n) \notin f_{cod}(c_i, w)$ (recall that $j = i$). Hence, (5) $\psi \notin f_{nor}(n)$. (3) and (5) imply that $\psi \notin f_{cod}^{c_i \boxplus n}(c_j)$ also when $j = i$, whence the result. □

So, as the procedure of code update we had before, the modified procedure of code update does not bring us outside the class of models under consideration, i.e., the suitable NC^+ models. The proof of the following theorem is now analogous to the proof of Theorem 7.3.8.

7.4.7. THEOREM. *The axiom system DNC^+, obtained from the axiom system DNC by replacing axioms $\mathsf{R}_{sat(c_j)}$ and $\mathsf{R}_{sat(c_i)}$ with the following axioms $\mathsf{R}'_{sat(c_j)}$ and $\mathsf{R}'_{sat(c_i)}$ is sound and complete with respect to the class of all suitable NC^+ models.*

$(\mathsf{R}'_{sat(c_j)})$ $[c_i \boxplus n] sat(c_j) \leftrightarrow (sat(c_j) \wedge \neg O_j \neg O_i sat(n))$, *for* $j \neq i$;

$(\mathsf{R}'_{sat(c_i)})$ $[c_i \boxplus n] sat(c_i) \leftrightarrow (sat(c_i) \wedge sat(n) \wedge \neg O_i \neg O_i sat(n))$.

The new axioms reflect the fact, after a code c_i is updated with a norm, the codes prescribing that agent i ought not to accept that norm are no longer satisfied.

7.5 Conclusion

In this chapter, we presented two explicit and dynamic deontic logics, DNC and its refinement DNC^+, for reasoning about the static and dynamic interaction between normative sources of two different kinds: norms, intended as elementary and consistent normative sources, and codes of agents, intended as complex and possibly inconsistent normative sources. The static component of the two systems allows us to represent several types of normative conflicts (i.e., local, global, between norms, between codes, between norms and codes) as well as the different senses in which an agent might accept a norm (i.e., simple acceptance of a norm *vs* acceptance of a norm as binding). The dynamic component of DNC and DNC^+ allows us to represent the procedure by means of which an agent accepts a norm as binding in a given situation. As shown by the examples of Antigone and Gandhi, dynamic operators make it possible to keep track, at least in simple situations, of which agent generates a normative conflict. This provides us with a first way to address an issue we left open in Chapter 6: Whose fault is it if a substandard state is reached? In the case of substandard situations resulting from the presence of a normative conflict, answering this question is also crucial to determine how the conflict is to be solved: conflicts generated in different ways, like those underlying cases of conscientious objection as opposed to cases of civil disobedience, may require different solutions.

Let us conclude by mentioning some directions for future work. From a purely logical point of view, one main question is how to generalize DNC^+ in a way that it allows obligation operators to nest arbitrarily and, relatedly, agents to update their codes with prescriptions concerning what other agents ought to do. The key problem is to track changes in the satisfaction of a code c_i deriving from the update of another code c_j to which c_i refers either directly (as in: $O_iO_j\varphi$) or indirectly (as in: $O_iO_{k_1}\dots O_{k_n}O_j\varphi$).

From a more conceptual point of view, in this chapter, we have focused on a simple kind of normative conflicts, namely conflicts arising between two agents and generated by expanding an agent's code with a norm. Yet, in daily life, normative conflicts often emerge within groups of more than two agents and, possibly, as a result of more complex changes to the agents' codes than merely "accepting a norm." Concerning the first issue, representing conflicts among more than two agents in DNC^+ is, by itself, not problematic: the definitions of global and local conflict between two agents [Definition 7.2.2] can be easily generalized to any (finite) number of agents. Since we take agents to include not only individual human beings but also organizations, communities, states, etc. [cf. footnote 1], a more interesting line of research would be to study how conflicts transfer and possibly spread from a layer to another of an organizational structure. This, of course, would require us to represent the relations between agents belonging to such a structure, as it is done, e.g., in Grossi et al. [2004, 2005, 2007]. Concerning the second issue, the key point is that, in principle, conflicts might arise from

procedures of code update that are akin to revision and contraction (rather than expansion) in the AGM theory [Alchourrón et al., 1985]. In the case of *legal* codes, even more complex procedures referring to the so-called "time of a norm" would be required in order to model changes like legal abrogation or annulments [see, e.g., Governatori and Rotolo, 2010]. We conjecture that techniques from DEL designed to model phenomena like forgetting [see, e.g., Fernández–Duque et al., 2015] may help in improving our dynamic deontic framework in this direction.

Appendix A

Appendix of Chapter 3

A.1 Completeness of ALO_n

In this appendix we prove Theorem 3.2.8. The proof that the axiom system ALO_n is sound with respect to the class of all ALO_n frames is a matter of routine validity check and it is thus omitted. The proof of completeness consists of two main steps. First, we define a Kripke semantics for $\mathcal{L}_{\mathsf{ALO}_n}$ and prove that ALO_n is sound and complete with respect to the class of Kripke models for $\mathcal{L}_{\mathsf{ALO}_n}$ (called *pseudo-models*). We then show that every formula of $\mathcal{L}_{\mathsf{ALO}_n}$ that is satisfiable in a pseudo-model is also satisfiable in an ALO_n model. By adapting the technique presented in Herzig and Lorini [2010], we do this by showing that every pseudo-model in which a formula $\varphi \in \mathcal{L}_{\mathsf{ALO}_n}$ is satisfiable can be turned into an ALO_n model in which φ is satisfiable. The intuitive ideas underlying the two steps of the proof are the same as those discussed in Chapter 2.2.1 and Chapter 2.2.3. We will adopt the terminology introduced in those sections and keep intuitive explanations to a minimum in what follows.

A.1.1 Kripke semantics for $\mathcal{L}_{\mathsf{ALO}_n}$

With respect to Kripke models for temporal STIT [see Chapter 2.2.3], pseudo-models are built from frames featuring a relation R_{X} that represents what happens next, a function f_{do} that assigns to every possible state the complete group action that is performed at that state, and a function $f_{\triangleright}$ that assigns to every (individual and group) action the set of (individual and group) actions that oppose it. As in Kripke models for temporal STIT, the relation $R_{\Box}$ represents the moment-partition. As usual, for any binary relation R on a set S and any $s \in S$, we define $R(s) = \{s' \in S \mid sRs'\}$.

A.1.1. Definition (Kripke ALO_n frame). A Kripke ALO_n frame is a tuple

$$\langle W, R_{\Box}, R_{\mathsf{X}}, f_{do}, f_{\triangleright} \rangle$$

where $W \neq \varnothing$ is a set of possible states, $R_\Box \subseteq W \times W$ is an equivalence relation, $R_\mathsf{X} \subseteq W \times W$ is the *next relation*, $f_{do} : W \to \mathit{Ag}\text{-}\mathit{Acts}$ is the *action function*, and $f_\triangleright : \mathit{G}\text{-}\mathit{Acts} \to 2^{\mathit{G}\text{-}\mathit{Acts}}$ is the *opposing function*. For any $w \in W$ and $i \in \mathit{Ag}$, let:

$\mathit{Acts}_i^w = \{f_{do}(w')(i) \in \mathit{Acts}_i \mid w' \in R_\Box(w)\}$ be the *actions available to* i *at* $R_\Box(w)$;

$\mathit{Acts}^w = \bigcup_{i \in \mathit{Ag}} \mathit{Acts}_i^w$ be the *individual actions executable at* $R_\Box(w)$.

In addition, define $R_{\mathit{Ag}} \subseteq W \times W$ by setting: for all $w, w' \in W$,

$wR_{\mathit{Ag}}w'$ iff $wR_\Box w'$ and $f_{do}(w) = f_{do}(w')$.

The elements of Kripke ALO_n frames satisfy the following conditions:

1. *Properties of* R_X: for all $w, w_1, w_2 \in W$,
   Seriality: there is $w' \in W$ such that $wR_\mathsf{X}w'$.
   Functionality: if $wR_\mathsf{X}w_1$ and $wR_\mathsf{X}w_2$, then $w_1 = w_2$.
2. *Independence of Agents*: for all $w \in W$ and $\alpha \in \mathit{Ag}\text{-}\mathit{Acts}$, if $\alpha(j) \in \mathit{Acts}^w$ for all $j \in \mathit{Ag}$, then there is $w' \in R_\Box(w)$ such that $f_{do}(w') = \alpha$.
3. *No Choice between Undivided Histories*: for all $w_1, w_2, w_3 \in W$, if $w_1R_\mathsf{X}w_2$ and $w_2R_\Box w_3$, then there is $v \in W$ such that $w_1R_{\mathit{Ag}}v$ and $vR_\mathsf{X}w_3$.
4. *Properties of* $f_\triangleright$: for all $\alpha_I, \beta_J, \gamma_K \in \mathit{G}\text{-}\mathit{Acts}$,
   Irreflexivity of opposing: $\alpha_I \notin \mathbf{opp}(\alpha_I)$.
   Monotonicity of Opposing: if $\alpha_I \in f_\triangleright(\beta_J)$ and $\beta_J \sqsubseteq \gamma_K$, then $\alpha_I \in f_\triangleright(\gamma_K)$.

We now define pseudo-models based on Kripke ALO_n frames and truth for formulas from $\mathcal{L}_{\mathsf{ALD}_n}$ at a state.

A.1.2. Definition (Pseudo-model). A pseudo-model is a tuple $\langle F, \nu \rangle$, where F is a Kripke ALO_n frame and $\nu : \mathit{Prop} \to 2^W$ is a valuation function.

A.1.3. Definition (Kripke semantics for $\mathcal{L}_{\mathsf{ALO}_n}$). Given a pseudo-model M, truth of a formula $\varphi \in \mathcal{L}_{\mathsf{ALO}_n}$ at a state w in M, denoted $M, w \models \varphi$, is defined recursively. Truth of atomic propositions and the Boolean connectives is defined as usual. The remaining clauses are as follows:

$$
\begin{array}{ll}
M, w \models \mathit{do}(a_i) & \text{iff } f_{do}(w)(i) = a_i \\
M, w \models \alpha_I \triangleright \beta_J & \text{iff } \alpha_I \in f_\triangleright(\beta_J) \\
M, w \models \mathsf{X}\varphi & \text{iff for all } w' \in W, \text{ if } wR_\mathsf{X}w', \text{ then } M, w' \models \varphi \\
M, w \models \Box\varphi & \text{iff for all } w' \in W, \text{ if } wR_\Box w', \text{ then } M, w' \models \varphi
\end{array}
$$

A.1.4. Theorem. *The axiom system* ALO_n*, defined by the axioms and rules in Table 3.2, is sound and complete with respect to the class of Kripke* ALO_n *frames.*

The proof of soundness is proceeds as usual and it is thus omitted. The proof of completeness is based on the construction of a canonical model for ALO_n. For any set of formulas $\Gamma \subseteq \mathcal{L}_{\mathsf{ALO}_n}$ and formula $\varphi \in \mathcal{L}_{\mathsf{ALO}_n}$, we write $\vdash_{\mathsf{ALO}_n} \varphi$ if φ is a theorem of ALO_n and $\Gamma \vdash_{\mathsf{ALO}_n} \varphi$ if there are finitely many formulas $\psi_1, \ldots, \psi_n \in \Gamma$ such that $\vdash_{\mathsf{ALO}_n} \psi_1 \wedge \cdots \wedge \psi_n \to \varphi$. We say that Γ is ALO_n-*consistent* if $\Gamma \nvdash_{\mathsf{ALO}_n} \bot$ and ALO_n-*inconsistent* otherwise. A formula $\varphi \in \mathcal{L}_{\mathsf{ALO}_n}$ is ALO_n-*consistent with* Γ if $\Gamma \cup \{\varphi\}$ is consistent (equivalently, if $\Gamma \nvdash_{\mathsf{ALO}_n} \neg\varphi$) and ALO_n-*consistent* if it is consistent with $\varnothing$. Finally, a set of formulas is a *maximally* ALO_n-*consistent set* (henceforth: mcs) if it is ALO_n-consistent and any proper superset of Γ is ALO_n-inconsistent. We omit reference to the logic ALO_n when it is clear from the context. The next Lemmas A.1.5 and A.1.6 are standard results about mcs [see Blackburn et al., 2001, Chapter 4.2], which we will repeatedly use in the proofs below without explicit mention.

A.1.5. Lemma. *For every mcs* w *of* ALO_n *and* $\varphi, \psi \in \mathcal{L}_{\mathsf{ALO}_n}$*, the following hold:*

1. $w \vdash_{\mathsf{ALO}_n} \varphi$ *iff* $\varphi \in w$,
2. *if* $\varphi \in w$ *and* $\varphi \to \psi \in w$*, then* $\psi \in w$,
3. $\neg\varphi \in w$ *iff* $\varphi \notin w$
4. $\varphi \wedge \psi \in w$ *iff* $\varphi \in w$ *and* $\psi \in w$.

A.1.6. Lemma (Lindebaum's Lemma). *Every maximal consistent set can be extended to a mcs.*

Let $\mathcal{W}$ be the set of all mcs of ALO_n and, for any $w \in \mathcal{W}$, let

1. $w/\blacksquare = \{\varphi \in \mathcal{L}_{\mathsf{ALO}_n} \mid \blacksquare\varphi \in w\}$, where $\blacksquare \in \{\Box, \mathsf{X}\}$;
2. $pos_{\triangleright}(w) = \{\alpha_I \triangleright \beta_J \in \mathcal{L}_{\mathsf{ALO}_n} \mid \alpha_I \triangleright \beta_J \in w\}$;
3. $neg_{\triangleright}(w) = \{\neg(\alpha_I \triangleright \beta_J) \in \mathcal{L}_{\mathsf{ALO}_n} \mid \neg(\alpha_I \triangleright \beta_J) \in w\}$.

A.1.7. Definition (Canonical ALO_n model for w_0). Let w_0 be a mcs. The canonical ALO_n model for w_0 is a tuple $M^c = \langle W^c, R^c_{\Box}, R^c_{\mathsf{X}}, f^c_{do}, f^c_{\triangleright}, \nu^c \rangle$, where

- $W^c = \{w \in \mathcal{W} \mid pos_{\triangleright}(w_0) \cup neg_{\triangleright}(w_0) \subseteq w\}$;
- $R^c_{\Box} \subseteq W^c \times W^c$ is such that, for all $w, w' \in W^c$, $wR^c_{\Box}w'$ iff $w/\Box \subseteq w'$;
- $R^c_{\mathsf{X}} \subseteq W^c \times W^c$ is such that, for all $w, w' \in W^c$, $wR^c_{\mathsf{X}}w'$ iff $w/\mathsf{X} \subseteq w'$;
- $f^c_{do} : W^c \to \mathit{Ag\text{-}Acts}$ is such that, for all $w \in W^c$, $f^c_{do}(w) = \alpha$ iff $do(\alpha) \in w$;
- $f^c_{\triangleright} : \mathit{G\text{-}Acts} \to 2^{\mathit{G\text{-}Acts}}$ is such that, for all $\alpha_I, \beta_J \in \mathit{G\text{-}Acts}$, $\alpha_I \in f^c_{\triangleright}(\beta_J)$ iff $\alpha_I \triangleright \beta_J \in w_0$;

- $\nu^c : Prop \to 2^{W^c}$ is such that, for all $w \in W^c$, $w \in \nu^c(p)$ iff $p \in w$.

The proof of the following Lemma A.1.8 relies on the next two theorems of ALO_n, which are an immediate consequence of $\mathsf{S5}_\Box$, KD_X, and of axioms Fun_X, $\mathsf{Sett}_\triangleright$ and $\mathsf{Fix}_\triangleright$:

(Thm1) $\quad \neg(\alpha_I \triangleright \beta_J) \to \Box\neg(\alpha_I \triangleright \beta_J)$.

(Thm2) $\quad \neg(\alpha_I \triangleright \beta_J) \to \mathsf{X}\neg(\alpha_I \triangleright \beta_J)$.

A.1.8. Lemma (Existence Lemma). *For all $w \in W^c$ and $\varphi \in \mathcal{L}_{\mathsf{ALO}_n}$, (a) if $\Diamond\varphi \in w$, then there is $w' \in W^c$ s.t. $w/\Box \subseteq w'$ and $\varphi \in w'$, and (b) if $\mathsf{X}\varphi \in w$, then there is $w' \in W^c$ s.t. $w/\mathsf{X} \subseteq w'$ and $\varphi \in w'$.*

Proof:
We only prove claim (a).[1] Let $\Diamond\varphi \in w$ and suppose, toward contradiction, that the set $\Gamma = pos_\triangleright(w_0) \cup neg_\triangleright(w_0) \cup w/\Box \cup \{\varphi\}$ is inconsistent. Then, there are finite sets $\Delta_1 \subseteq pos_\triangleright(w_0)$, $\Delta_2 \subseteq neg_\triangleright(w_0)$, and $\Delta_3 \subseteq w/\Box$ s.t. $\vdash_{\mathsf{ALO}_n} \bigwedge\Delta_1 \wedge \bigwedge\Delta_2 \wedge \bigwedge\Delta_3 \to \neg\varphi$. Hence, by the logic of $\Box$,

(*) $\vdash_{\mathsf{ALO}_n} \bigwedge_{\psi\in\Delta_1} \Box\psi \wedge \bigwedge_{\chi\in\Delta_2} \Box\chi \wedge \bigwedge_{\xi\in\Delta_3} \Box\xi \to \Box\neg\varphi$.

Since $\Delta_3 \subseteq w/\Box$, (1) $\bigwedge_{\xi\in\Delta_3} \Box\xi \in w$ by the def. of $w/\Box$. In addition, every $\psi \in \Delta_1$ has form $\alpha_I \triangleright \beta_J$ for some $\alpha_I, \beta_J \in$ *G-Acts*. By the def. of W^c, if $\alpha_I \triangleright \beta_J \in w_0$, then $\alpha_I \triangleright \beta_J \in w$, and so $\Box(\alpha_I \triangleright \beta_J) \in w$ by axiom $\mathsf{Sett}_\triangleright$. Hence, (2) $\Box\psi \in w$ for every $\psi \in \Delta_1$. Similarly, every $\chi \in \Delta_2$ has form $\neg(\alpha_I \triangleright \beta_J)$ for some $\alpha_I, \beta_J \in$ *G-Acts*. By the def. of W^c, if $\neg(\alpha_I \triangleright \beta_J) \in w_0$, then $\neg(\alpha_I \triangleright \beta_J) \in w$, and so $\Box\neg(\alpha_I \triangleright \beta_J) \in w$ by theorem Thm1. Hence, (3) $\Box\chi \in w$ for every $\chi \in \Delta_3$. It follows from (*), (1), (2), and (3) that $\Box\neg\varphi \in w$, and so $\Diamond\varphi \notin w$ against the hypothesis. Therefore, Γ is consistent and can be extended to a mcs w'. Since $pos_\triangleright(w_0) \cup neg_\triangleright(w_0) \subseteq w'$, $w' \in W^c$. In addition, $w/\Box \subseteq w'$. □

Given Lemma A.1.8, the proof of the next lemma follows the usual pattern [see Blackburn et al., 2001, Lem. 4.21, p. 201].

A.1.9. Lemma (Truth Lemma). *For all $w \in W^c$ and $\varphi \in \mathcal{L}_{\mathsf{ALO}_n}$,*

$$M^c, w \models \varphi \textit{ iff } \varphi \in w$$

A.1.10. Lemma. *The canonical ALO_n model M^c is a pseudo-model.*

[1] The proof of claim (b) proceeds as the proof of claim (a) except that axiom $\mathsf{Fix}_\triangleright$ and theorem Thm2 are used instead of axiom $\mathsf{Sett}_\triangleright$ and theorem Thm1.

Proof:
Since ALO_n includes the axioms of $\mathsf{S5}$ for $\Box$, the axioms of KD for X, and the axiom $\mathsf{Fun}_{\mathsf{X}}$, $R^c_{\Box}$ is an equivalence relation and R^c_{X} is a serial and functional relation by standard results in modal logic. In addition, it is immediate to check that axioms Act and Sin ensure that f_{do} is well-defined, while axioms $\mathsf{Mon}_{\triangleright}$ and $\mathsf{Irr}_{\triangleright}$ ensure that M^c satisfies the properties of irreflexivity and monotonicity of opposing. The proof that M^c satisfies the remaining properties of pseudo-models is as follows.

1. *M^c satisfies the condition of independence of agents.*

   Consider a mcs $w \in W^c$ and an action $\alpha \in \mathit{Ag\text{-}Acts}$ s.t., for all $j \in Ag$, there is $v_j \in W^c$ s.t. (1) $v_j \in R^c_{\Box}(w)$ and $f^c_{do}(v_j)(j) = \alpha(j)$. We need to prove that there is $w' \in R^c_{\Box}(w)$ s.t. $f^c_{do}(w') = \alpha$. Take any of the v_j satisfying (1). Since $f^c_{do}(v_j)(j) = \alpha(j)$, $do(\alpha(j)) \in v_j$ by the def. f^c_{do}. In addition, since $v_j \in R^c_{\Box}(w)$, $\Diamond do(\alpha(j)) \in w$ by Lem. A.1.9. Since, for every $j \in Ag$, there is a v_j satisfying (1), it follows that $\bigwedge_{j\in Ag} \Diamond do(\alpha(j)) \in w$. But then $\Diamond \bigwedge_{j\in Ag} do(\alpha(j)) \in w$ by axiom IA_{do}. It follows by Lem. A.1.8 that there is $w' \in W^c$ s.t. $w/\Box \subseteq w'$ and $\bigwedge_{j\in Ag} do(\alpha(j)) \in w'$. Hence, by the def. of $w/\Box$ and f^c_{do}, $w' \in R_{\Box}(w)$ and $f^c_{do}(w')(j) = \alpha(j)$ for all $j \in Ag$. That is, $f^c_{do}(w') = \alpha$, whence the result.

2. *M^c satisfies the condition of no choice between undivided histories.*

   Consider mcs $w_1, w_2, w_3 \in W^c$ s.t. (1) $w_1 R^c_{\mathsf{X}} w_2$ and (2) $w_2 R^c_{\Box} w_3$. By axiom Act, there is $\alpha \in \mathit{Ag\text{-}Acts}$ s.t. (3) $do(\alpha) \in w_1$. Hence, $f^c_{do}(w_1) = \alpha$ by the def. of f^c_{do}. We have to show that there is $v \in W^c$ s.t.

   (*i*) $w_1 R^c_{\Box} v$ (equivalently: $w_1/\Box \subseteq v$);

   (*ii*) $f^c_{do}(v) = \alpha$ (equivalently: $do(\alpha) \in v$);

   (*iii*) $v R^c_{\mathsf{X}} w_3$ (equivalently: $v/\mathsf{X} \subseteq w_3$).

   It is an easy exercise to show that $v/\mathsf{X} \subseteq w_3$ iff $\{\mathsf{X}\varphi \in \mathcal{L}_{\mathsf{ALO}_n} \mid \varphi \in w_3\} \subseteq v$. So, consider any $\varphi \in w_3$. Given hypotheses (1) and (2), $\hat{\mathsf{X}}\Diamond\varphi \in w_1$ by Lem. A.1.9 and Def. A.1.3. So, $do(\alpha) \wedge \mathsf{X}\Diamond\varphi \in w_1$ by (3) and axiom $\mathsf{Fun}_{\mathsf{X}}$. Hence, $\Diamond(do(\alpha) \wedge \mathsf{X}\varphi) \in w_1$ by axiom UH_{do}. But then, there is $v \in W^c$ s.t. $w_1/\Box \subseteq v$ and $do(\alpha) \wedge \mathsf{X}\varphi \in v$ by Lem. A.1.8. Therefore, (*i*) $w_1/\Box \subseteq v$, (*ii*) $do(\alpha) \in v$, and, since φ was an arbitrary formula in w_3, (*iii*) $\{\mathsf{X}\varphi \in \mathcal{L}_{\mathsf{ALO}_n} \mid \varphi \in w_3\} \subseteq v$, as it was desired.

□

The previous lemmas suffice to conclude that ALO_n is complete with respect to the class of Kripke ALO_n frames. To see this, let $\varphi \in \mathcal{L}_{\mathsf{ALO}_n}$ be such that

$\nvdash_{\mathsf{ALO}_n} \varphi$. Then, $\{\neg\varphi\}$ is consistent, and so it can be extended to a mcs w_0 by Lindenbaum's Lemma [Lem. A.1.6]. Let $M^c = \langle W^c, R^c_\Box, R^c_{\mathsf{X}}, f^c_{do}, f^c_{\triangleright}, \nu^c\rangle$ be the canonical ALO_n model for w_0. Since $pos_{\triangleright}(w_0) \cup neg_{\triangleright}(w_0) \subseteq w_0$ by the def. of $pos_{\triangleright}(w_0)$ and $neg_{\triangleright}(w_0)$, $w_0 \in W^c$. In addition, by the properties of mcs [Lem. A.1.5], $\varphi \notin w_0$. Therefore, $M^c, w_0 \nvDash \varphi$ by the Truth Lemma [Lem. A.1.9]. Since M^c is a pseudo-model [Lem. A.1.10], we conclude that φ is not valid in the class of Kripke ALO_n frames.

A.1.2 From pseudo-models to ALO_n models

Let φ_0 be an ALO_n-consistent formula of $\mathcal{L}_{\mathsf{ALO}_n}$. Then, by Theorem A.1.4 there is a pseudo-model $M = \langle W, R_\Box, R_{\mathsf{X}}, f_{do}, f_{\triangleright}, \nu\rangle$ such that $M, \underline{w} \models \varphi_0$ for some $\underline{w} \in W$. We want to show that M can be transformed into an ALO_n model in which φ_0 is satisfiable. The construction requires two preliminary steps. First, we unravel M [see Blackburn et al., 2001, p. 63] in order to ensure that the relation R_{X} generates an acyclic ordering on the equivalence classes in the moment-partition. Second, we force every cell in the moment-partition that is far enough from $\underline{w}$ along the relation R_{X} to be a singleton. This will ensure that every state in the unraveled model will correspond to one and only one index in the ALO_n model built from it.

First preliminary step

A.1.11. Definition (Unraveled model M'). Define $M' = \langle W', R'_\Box, R'_{\mathsf{X}}, f'_{do}, f'_{\triangleright}, \nu\rangle$ so that:

- W' is the set of all sequences $\overrightarrow{w_n} = w_1 w_2 \dots w_n$ such that $w_1 R_\Box \underline{w}$ and, for all $1 \leq i < n$ (with $n \in \mathbb{N}$), $w_i R_{\mathsf{X}} w_{i+1}$;
- $R'_\Box \subseteq W' \times W'$ is such that, for all $\overrightarrow{w_n}, \overrightarrow{v_m} \in W'$, $\overrightarrow{w_n} R'_\Box \overrightarrow{v_m}$ iff $n = m$, $w_i R_\Box v_i$ for all $1 \leq i \leq n$, and $f_{do}(w_i) = f_{do}(v_i)$ for all $1 \leq i < n$;
- $R'_{\mathsf{X}} \subseteq W' \times W'$ is such that, for all $\overrightarrow{w_n}, \overrightarrow{v_m} \in W'$, $\overrightarrow{w_n} R'_{\mathsf{X}} \overrightarrow{v_m}$ iff $\overrightarrow{v_m} = \overrightarrow{w_n} v_m$ and $w_n R_{\mathsf{X}} v_m$;
- $f'_{do} : W' \to \mathit{Ag\text{-}Acts}$ is such that, for all $\overrightarrow{w_n} \in W'$, $f'_{do}(\overrightarrow{w_n}) = f_{do}(w_n)$;
- $f'_{\triangleright} : \mathit{G\text{-}Acts} \to 2^{\mathit{G\text{-}Acts}}$ is such that, for all $\alpha_I \in \mathit{G\text{-}Acts}$, $f'_{\triangleright}(\alpha_I) = f_{\triangleright}(\alpha_I)$;
- $\nu' : \mathit{Prop} \to 2^{W'}$ is such that, for all $\overrightarrow{w_n} \in W'$, $\overrightarrow{w_n} \in \nu'(p)$ iff $w_n \in \nu(p)$

The following proposition highlights a key fact about the unraveled model M'.

A.1.12. Proposition. *For all $\overrightarrow{w_n} \in W'$ and $v \in W$, if $w_n R_\Box v$, then there is $\overrightarrow{v_n} \in W'$ such that $v_n = v$ and $\overrightarrow{w_n} R'_\Box \overrightarrow{v_n}$.*

Proof:
The proof is by induction on n. The case for $n = 1$ is trivial. So, assume that the claim holds for n and consider a sequence $\overrightarrow{w_{n+1}} \in W'$ such that there is $v \in W$ with (1) $w_{n+1}R_{\Box}v$. By the def. of W', (2) $w_nR_{\mathsf{X}}w_{n+1}$. Since M is a pseudo model, (1) and (2) entail that there is $u \in W$ s.t. $w_nR_{Ag}u$ and $uR_{\mathsf{X}}v$ by the condition of no choice between undivided histories. By the def. of R_{Ag}, (3) $w_nR_{\Box}u$ and (4) $f_{do}(w_n) = f_{do}(u)$. By the inductive hypothesis, it follows from (3) that there is $\overrightarrow{u_n} \in W'$ s.t. $u_n = u$ and (5) $\overrightarrow{w_n}R'_{\Box}\overrightarrow{u_n}$. Given the def. of $R'_{\Box}$, (1), (4), and (5) suffice to conclude that $\overrightarrow{w_{n+1}}R'_{\Box}\overrightarrow{u_n}v$. □

A.1.13. Proposition. *The unraveled model M' is a pseudo-model.*

Proof:
Since M is a pseudo model, it is immediate to check that Definition A.1.11 ensures that $R'_{\Box}$ is an equivalence relation, that R'_{X} is serial and functional, and that $f'_{\triangleright}$ satisfies the properties of irreflexivity and monotonicity of opposing. Let us consider the remaining conditions.

1. *M' satisfies the condition of independence of agents.*

   Consider any $\overrightarrow{w_n} \in W'$ and $\alpha \in Ag\text{-}Acts$ s.t., for all $j \in Ag$, there is $\overrightarrow{v_{n_j}} \in R'_{\Box}(\overrightarrow{w_n})$ s.t. $f'_{do}(\overrightarrow{v_{n_j}}) = \alpha(j)$. Then, by the def. of $R'_{\Box}$ and f'_{do}, for all $j \in Ag$, there is v_{n_j} s.t. $v_{n_j} \in R_{\Box}(w_n)$ and $f_{do}(v_{n_j}) = \alpha(j)$. Since M is a pseudo-model, it follows that there is $u \in R_{\Box}(w_n)$ s.t. $f_{do}(u) = \alpha$ by the condition of independence of agents. By applying Prop. A.1.12 and the def. of f'_{do}, we conclude that there is $\overrightarrow{u_n} \in W'$ s.t. $\overrightarrow{u_n} \in R'_{\Box}(\overrightarrow{w_n})$ and $f'_{do}(\overrightarrow{u_n}) = f_{do}(u) = \alpha$.

2. *M' satisfies the condition of no choice between undivided histories.*

   Consider $\overrightarrow{w_n}, \overrightarrow{v_m}, \overrightarrow{u_m} \in W'$ s.t. (1) $\overrightarrow{w_n}R'_{\mathsf{X}}\overrightarrow{v_m}$ and (2) $\overrightarrow{v_m}R'_{\Box}\overrightarrow{u_m}$. By the def. of R'_{X} and $R'_{\Box}$, $w_nR_{\mathsf{X}}v_m$ and $v_mR_{\Box}u_m$. Since M is a pseudo-model, it follows that there is $x \in W$ s.t. $w_nR_{Ag}x$ (i.e., (3) $w_nR_{\Box}x$ and (4) $f_{do}(w_n) = f_{do}(x)$) and (5) $xR_{\mathsf{X}}u_m$ by the condition of no choice between undivided histories. Since $\overrightarrow{w_n} \in W'$, from (3) and Prop. A.1.12 it follows that there is $\overrightarrow{x_n} \in W'$ s.t. $x_n = x$ and (6) $\overrightarrow{w_n}R'_{\Box}\overrightarrow{x_n}$. By (4) and the def. of f'_{do}, (7) $f'_{do}(\overrightarrow{x_n}) = f'_{do}(\overrightarrow{w_n})$. (6) and (7) entail that $\overrightarrow{w_n}R'_{Ag}\overrightarrow{x_n}$. Finally, it follows from (5) that $\overrightarrow{x_n}R'_{\mathsf{X}}\overrightarrow{u_m}$, whence the result.

□

A.1.14. Proposition. *Let $f : W' \to W$ be the mapping defined by setting, for all $\overrightarrow{w_n} \in W'$: $f(\overrightarrow{w_n}) = w_n$. Then, for all $\overrightarrow{w_n} \in W'$ and $\varphi \in \mathcal{L}_{\mathsf{ALO}_n}$, $M', \overrightarrow{w_n} \models \varphi$ iff $M, f(\overrightarrow{w_n}) \models \varphi$.*

Proof:
By standard results in modal logic [Blackburn et al., 2001, Prop. 2.14, p. 62], it suffices to show that the function f is a bounded morphism from M' to M. That is, letting $R_{\blacksquare} \in \{R_{\Box}, R_{\mathsf{X}}\}$, we need to check that the following are satisfied for all $\overrightarrow{w_n}, \overrightarrow{v_m} \in W'$, $\alpha_I \in \mathit{G\text{-}Acts}$, and $p \in \mathit{Prop}$:

1. if $\overrightarrow{w_n} R'_{\blacksquare} \overrightarrow{v_m}$, then $f(\overrightarrow{w_n}) R_{\blacksquare} f(\overrightarrow{v_m})$;
2. if $f(\overrightarrow{w_n}) R_{\blacksquare} v$, then there is $\overrightarrow{v_m} \in W'$ s.t. $f(\overrightarrow{v_m}) = v$ and $\overrightarrow{w_n} R'_{\blacksquare} \overrightarrow{v_m}$;
3. $f'_{do}(\overrightarrow{w_n}) = f_{do}(f(\overrightarrow{w_n}))$;
4. $f'_{\triangleright}(\alpha_I) = f_{\triangleright}(\alpha_I)$;
5. $\overrightarrow{w_n} \in \nu'(p)$ iff $f(\overrightarrow{w_n}) \in \nu(p)$

The proof follows immediately from Def. A.1.11 in all cases, except for case 2 when $\blacksquare$ is $\Box$. But this case coincides with Prop. A.1.12. □

A.1.15. Corollary. *Let $\overrightarrow{\underline{w}}$ be the one-element sequence consisting of $\underline{w}$. Then, $M', \overrightarrow{\underline{w}} \models \varphi_0$.*

Second preliminary step

We now want to turn M' into an ALO_n model. The idea is simple [cf. Chapter 2.2.1 and Chapter 2.2.3]: we take equivalence classes determined by $R'_{\Box}$ as moments, and we show that R'_{X} induces a tree-like ordering on moments. But before doing this we need to take an extra step to ensure that states in the pseudo-model M' and moment-history pairs in the ALO_n model built from M' are in a one-to-one correspondence. This will allow us to define a valuation function for the generated ALO_n model from the valuation ν' in a straightforward way.

A.1.16. Definition. (Unraveled model M'') Let x be the modal X-depth of φ_0, i.e., the maximum number of nested X modalities in φ_0. M'' is the tuple obtained from M' by replacing $R'_{\Box}$ with the relation $R''_{\Box} \subseteq W' \times W'$ defined by setting, for all $\overrightarrow{w_n}, \overrightarrow{v_m} \in W'$:

$$\overrightarrow{w_n} R''_{\Box} \overrightarrow{v_m} \text{ iff } \begin{cases} \overrightarrow{w_n} R'_{\Box} \overrightarrow{v_m} & \text{if } n \leq \mathsf{x} \\ \overrightarrow{w_n} = \overrightarrow{v_m} & \text{if } \mathsf{x} < n \end{cases}$$

So, in M'', all sequences of length $n > \mathsf{x}$ belong to a singleton equivalence class of $R''_{\Box}$. It is immediate to check that M'' is still a pseudo-model. In addition, the next proposition follows straightforwardly from Corollary A.1.15 and from the fact that states in W' that are equivalent under $R'_{\Box}$ are separated only at a R'_{X}-depth that is higher than the modal X-depth of φ_0.

A.1.17. Proposition. $M'', \underline{w} \models \varphi_0$.

From M'' to a ALO_n model

We start by building a DBT structure from M''. We write $W'/R''_\Box$ for the quotient set of W' by $R''_\Box$ and, for any $\overrightarrow{w_n} \in W'$, we write $[\overrightarrow{w_n}]$ for the equivalence class of $\overrightarrow{w_n}$ in $W'/R''_\Box$.

A.1.18. Definition (Generated DBT structure). Let $\mathcal{T} = \langle Mom, < \rangle$, where

- $Mom = W'/R''_\Box$;
- $< \subseteq Mom \times Mom$ is s.t., for all $[\overrightarrow{w_n}], [\overrightarrow{v_m}] \in Mom$,

$$[\overrightarrow{w_n}] < [\overrightarrow{v_m}] \text{ iff } n < m \text{ and, for all } \overrightarrow{u_m} \in [\overrightarrow{v_m}], \overrightarrow{u_n} \in [\overrightarrow{w_n}]$$

(i.e., $[\overrightarrow{w_n}] < [\overrightarrow{v_m}]$ iff all prefixes of length n of sequences in $[\overrightarrow{v_m}]$ are in $[\overrightarrow{w_n}]$).

A.1.19. Proposition. *$\mathcal{T}$ is a DBT structure.*

Proof:
Since, for all $\overrightarrow{w_n}, \overrightarrow{v_m} \in W''$, $\overrightarrow{w_n} R''_\Box \overrightarrow{v_m}$ only if $n = m$ by the def. of $R''_\Box$, the relation $<$ is well-defined. That $<$ is irreflexive, transitive, and discrete follows immediately from the definition. Seriality follows from the fact that R'_{X} is serial. Finally, for past-linearity consider $[\overrightarrow{w_n}], [\overrightarrow{v_m}], [\overrightarrow{u_k}] \in Mom$ s.t. (1) $[\overrightarrow{w_n}] \leq [\overrightarrow{u_k}]$ and (2) $[\overrightarrow{v_m}] \leq [\overrightarrow{u_k}]$, and suppose, toward contradiction, that (3) $[\overrightarrow{w_n}] \not\leq [\overrightarrow{v_m}]$ and (4) $[\overrightarrow{v_m}] \not\leq [\overrightarrow{w_n}]$. Either $n \leq m$ or $m \leq n$. Suppose the former is the case. Given (3), we can suppose, without any loss in generality, that (*) it is not the case that $\overrightarrow{w_n} R''_\Box \overrightarrow{v_n}$. Yet, from (1) and the def. of $<$, we get (5) $\overrightarrow{w_n} R''_\Box \overrightarrow{u_n}$ and, from (2) and the def. of $<$, we get (6) $\overrightarrow{u_m} R''_\Box \overrightarrow{v_m}$. Since $n \leq m$, (6) implies that $\overrightarrow{u_n} R''_\Box \overrightarrow{v_n}$ by the def. of $R''_\Box$. But this, together with (5), implies that $\overrightarrow{w_n} R''_\Box \overrightarrow{v_n}$, which contradicts (*). By reasoning in an analogous way, a contradiction is reached if $m \leq n$. Hence, either $[\overrightarrow{w_n}] \leq [\overrightarrow{v_m}]$ or $[\overrightarrow{v_m}] \leq [\overrightarrow{w_n}]$. □

The notions of history in $\mathcal{T}$, successor of a moment, index in $\mathcal{T}$, and the related notation are introduced in the usual way [see Table 3.1].

We now want to show that there is a one-to-one correspondence between possible states in the pseudo-model M'' and indices in $\mathcal{T}$. The key idea is that, for every index $[\overrightarrow{w_n}]/h \in Ind^{\mathcal{T}}$, history h has a "witnessing state" in $[\overrightarrow{w_n}]$ [see Chapter 2.2.3 for more on this]. To make this idea precise, let us highlight some important facts. Given the construction of M'', every history $h \in Hist^{\mathcal{T}}$ has a beginning, namely moment $[\overrightarrow{\underline{w}}]$. For any $n \in \mathbb{N}$ and $h \in Hist^{\mathcal{T}}$, define $h(n)$ inductively as follows:

1. $h(0) = [\overrightarrow{\underline{w}}]$;
2. $h(n+1) = succ_h(h(n))$.

Intuitively, $h(n)$ is the n-th moment on h. Any index $[\overrightarrow{w_n}]/h \in Ind^{\mathcal{T}}$ can then be re-written as $h(n)/h$. Recall that x is the modal X-depth of φ_0. Definition A.1.16 and the functionality of R'_{X} ensure that:

1. for all $n > \mathsf{x}$, $h(n)$ is a singleton – we write $\overrightarrow{w_{h(n)}}$ for its only element;
2. for all $n > \mathsf{x}$ and $h, h' \in Hist^{\mathcal{T}}$, if $h(n) = h'(n)$, then $h = h'$.

A.1.20. Definition. Let $\omega : Ind^{\mathcal{T}} \to W'$ be the mapping such that, for all $h(n)/h \in Ind^{\mathcal{T}}$, $\omega(h(n)/h)$ is the prefix of length n of $\overrightarrow{w_{h(n+\mathsf{x})}}$.

Intuitively, the function ω finds, for every index $h(n)/h$, the witness of h in $h(n)$. It does so by picking a singleton moment on h that occurs later than $h(n)$, and by selecting the prefix of length n of the unique element of this moment. Observe that the prefix in question belongs to $h(n)$ by the definition of $<$: that is, $\omega(h(n)/h) \in h(n)$ for all $h(n)/h \in Ind^{\mathcal{T}}$. It is an easy exercise to prove that ω is a bijection. We write ω^{-1} for its inverse. The following facts will be useful below.

A.1.21. Fact. Let $h, h_1, h_2 \in Hist^{\mathcal{T}}$ and $n \in \mathbb{N}$.

(*a*) If $h_1(n) = h_2(n)$, then h_1 and h_2 are undivided at $h_1(n)$ iff $h_1(n+1) = h_2(n+1)$;

(*b*) $h_1(n) = h_2(n)$ iff $\omega(h_1(n)/h_1)\, R''_{\Box}\, \omega(h_2(n)/h_2)$;

(*c*) $\omega(h(n)/h)\, R'_{\mathsf{X}}\, \omega(h(n+1)/h)$.

Proof:
(*a*) Immediate by the def. of $h(n)$. (*b*) Immediate, as $\omega(h(n)/h) \in h(n)$ for all $h(n)/h \in Ind^{\mathcal{T}}$. (*c*) By the def. of ω, $\omega(h(n+1)/h)$ is the prefix of length $n+1$ of $\overrightarrow{w_{h(n+1+\mathsf{x})}}$. Let $\mathrm{pref}(n)$ be the prefix of length n of this sequence. By the def. of $<$, $\mathrm{pref}(n) \in h(n+\mathsf{x})$. But $h(n+\mathsf{x})$ is a singleton. So $\mathrm{pref}(n)$ is the prefix of length n of $\overrightarrow{w_{h(n+\mathsf{x})}}$. Hence, $\mathrm{pref}(n) = \omega(h(n)/h)$ by the def. of ω. By the def. of R'_{X} it follows that $\omega(h(n)/h) R'_{\mathsf{X}} \omega(h(n+1)/h)$. □

A.1.22. Definition (Generated ALO_n model). Let $\mathcal{M} = \langle \mathcal{T}, \mathbf{act}, \mathbf{opp}, \pi \rangle$, where

- $\mathcal{T}$ is defined as in Definition A.1.18;
- $\mathbf{act} : Ind^{\mathcal{T}} \to Ag\text{-}Acts$ is such that, for all $h(n)/h \in Ind^{\mathcal{T}}$, $\mathbf{act}(h(n)/h) = f'_{do}(\omega(h(n)/h))$;
- $\mathbf{opp} : G\text{-}Acts \to 2^{G\text{-}Acts}$ is such that, for all $\alpha_I \in G\text{-}Acts$, $\mathbf{opp}(\alpha_I) = f_{\triangleright}(\alpha_I)$;

- $\pi : Prop \to 2^{Ind^{\mathcal{T}}}$ is such that, for all $p \in Prop$ and $h(n)/h \in Ind^{\mathcal{T}}$, $h(n)/h \in \pi(p)$ iff $\omega(h(n)/h) \in \nu'(p)$.

A.1.23. Proposition. *$\mathcal{M}$ is a ALO_n model.*

Proof:
We have already checked that $\mathcal{T}$ is a DBT structure. In addition, it follows immediately from its definition and the fact that M'' is a pseudo-model that **opp** satisfies the conditions of irreflexivity and monotonicity of opposing. We prove that $\mathcal{M}$ satisfies the remaining conditions.

1. *No choice between undivided histories.*

   Take any two histories $h_1, h_2 \in Hist^{\mathcal{T}}$ such that (1) $h_1(n) = h_2(n)$ and (2) $h_1(n+1) = h_2(n+1)$. We have to show that $\mathbf{act}(h_1(n)/h_1) = \mathbf{act}(h_2(n)/h_2)$. By Fact A.1.21(c), (3) $\omega(h_1(n)/h_1)\, R'_{\mathsf{X}}\, \omega(h_1(n+1)/h_1)$. In addition, it follows from (2) that (4) $\omega(h_1(n+1)/h_1)\, R''_{\Box}\, \omega(h_2(n+1)/h_2)$ by Fact A.1.21(c). Since M'' is a pseudo-model, (3) and (4) imply that there is $\overrightarrow{v_n} \in W'$ s.t. (5) $\omega(h_1(n+1)/h_1)\, R''_{Ag}\, \overrightarrow{v_n}$ and (6) $\overrightarrow{v_n}\, R'_{\mathsf{X}}\, \omega(h_2(n+1)/h_2)$ by the condition of no choice between undivided histories. Since $\omega(h_2(n)/h_2)\, R'_{\mathsf{X}}\, \omega(h_2(n+1)/h_2)$ by Fact A.1.21(c), $\overrightarrow{v_n} = \omega(h_2(n)/h_2)$ by the def. of R'_{X}. We can thus replace $\overrightarrow{v_n}$ with $\omega(h_2(n)/h_2)$ in (5) and obtain that $f_{do}(\omega(h_1(n+1)/h_1)) = f_{do}(\omega(h_2(n+1)/h_2))$. Hence, $\mathbf{act}(h_1(n)/h_1) = \mathbf{act}(h_2(n)/h_2)$ by the def. of **act**.

2. *Independence of agents.*

   Suppose that there is $[\overrightarrow{w_n}] \in Mom$ and $\alpha \in Ag\text{-}Acts$ s.t., for all $j \in Ag$, there is $h_j \in Hist^{\mathcal{T}}$ s.t. $h_j(n) = [\overrightarrow{w_n}]$ and $\mathbf{act}(h_j(n)/h_j)(j) = \alpha(j)$. We have to show that there is $h \in Hist^{\mathcal{T}}$ s.t. $h(n) = [\overrightarrow{w_n}]$ and $\mathbf{act}(h(n)/h) = \alpha$. Given the hypothesis and the def. of **act**, we know that, for all $j \in Ag$, $f'_{do}(\omega(h_j(n)/h_j))(j) = \alpha(j)$. In addition, by the def. of ω, $\omega(h_j(n)/h_j) \in [\overrightarrow{w_n}]$ for all $j \in Ag$. Since M'' is a pseudo model, it follows that there is $\overrightarrow{v_n} \in W'$ s.t. $f'_{do}(\overrightarrow{v_n}) = \alpha$ by the condition of independence of agents. Since ω is a bijection, there is $h \in Hist^{\mathcal{T}}$ s.t. $\overrightarrow{v_n} = \omega(h(n)/h)$, where $h(n) = [\overrightarrow{v_n}] = [\overrightarrow{w_n}]$. By the def. of **act**, we conclude that $\mathbf{act}(h(n)/h) = \alpha$.

□

A.1.24. Proposition. *For all $\varphi \in \mathcal{L}_{\mathsf{ALO}_n}$ and $h(n)/h \in Ind^{\mathcal{T}}$, $\mathcal{M}, h(n)/h \models \varphi$ iff $M'', \omega(h(n)/h) \models \varphi$.*

Proof:
The proof is by induction on the complexity of φ. The cases for propositional variables,

Boolean connective, formulas like $do(a_i)$ and $\alpha_I \triangleright \beta_J$ are straightforward given Def. A.1.22. We prove the remaining cases.

1. $\varphi := \mathsf{X}\psi$

$\mathcal{M}, h(n)/h \models \mathsf{X}\psi$ iff $\mathcal{M}, h(n+1)/h \models \psi$ (by Def. 3.2.7)
iff $M'', \omega(h(n+1)/h) \models \psi$ (induction hypothesis)
iff for all $\overrightarrow{v_m} \in W'$ s.t. $\omega(h(n)/h) R'_{\mathsf{X}} \overrightarrow{v_m}$, $M'', \overrightarrow{v_m} \models \psi$
(by Fact A.1.21(c) and functionality R'_{X})
iff $M'', \omega(h(n)/h) \models \mathsf{X}\psi$ (by Def. 3.2.7)

2. $\varphi := \Box\psi$

$\mathcal{M}, h(n)/h \models \Box\varphi$ iff for all $h' \in Hist^{\mathcal{T}}$ s.t. $h(n) = h'(h)$, $\mathcal{M}, h'(n+1)/h' \models \varphi$
(by Def. 3.2.7)
iff for all $h' \in Hist^{\mathcal{T}}$ s.t. $h(n) = h'(h)$, $M, \omega(h'(n)/h') \models \varphi$
(by induction hypothesis)
iff for all $\overrightarrow{v_n} \in W'$ s.t. $\omega(h(n)/h)\, R''_{\Box}\, \overrightarrow{v_n}$, $M, \overrightarrow{v_n} \models \varphi$
(by Fact A.1.21 and the fact that ω is a bijection)
iff $M, \omega(h(n)/h) \models \varphi$ (by Def. 3.2.7)

□

Since φ_0 is an arbitrary ALO_n-consistent formula, the following corollary suffices to conclude the proof that ALO_n is complete with respect to the class of ALO_n frames.

A.1.25. Corollary. *There is a* ALO_n *model in which* φ_0 *is satisfiable.*

Proof:
Since $M'', \underline{\overrightarrow{w}} \models \varphi_0$ [Cor. A.1.15], $\mathcal{M}, \omega^{-1}(\underline{\overrightarrow{w}}) \models \varphi_0$ by Prop. A.1.24. In addition, $\mathcal{M}$ is an ALO_n model by Prop. A.1.23, whence the result. □

A.2 Logical relations between responsibility operators

A.2.1. Proposition. *Where* $I \subseteq Ag$ *and* $\varphi \in \mathcal{L}_{\mathsf{ALO}_n}$*, the following are* valid *in the class of* ALO_n *frames:*

1. $[I\, dxstit]\varphi \rightarrow [I\, pres]\varphi$

2. $[I\, sres]\varphi \rightarrow [I\, pres]\varphi$

3. $[I\, res]\varphi \rightarrow [I\, pres]\varphi$

4. $([I\, dxstit]\varphi \wedge \bigvee_{\alpha \in Ag\text{-}Acts}(do(\alpha) \wedge \bigwedge_{K \subset I}(\neg[\alpha_K]\varphi))) \rightarrow [I\, res]\varphi$

5. $([I\, pres]\varphi \wedge \bigvee_{\alpha \in Ag\text{-}Acts} \Box(do(\alpha_I) \rightarrow \underline{do}(\alpha_I))) \rightarrow [I\, dxstit]\varphi$

Proof:
Let $\mathcal{M}$ be any ALO_n model and m/h any index in $\mathcal{M}$. We prove that the above implications are true at m/h in $\mathcal{M}$.

1. If $\mathcal{M}, m/h \models [I\, dxstit]\varphi$, then (1) $\mathcal{M}, m/h \models do(\alpha_I) \wedge [\alpha_I]\varphi$ where $\alpha_I = \mathbf{act}(m/h)_I$ and (2) $\mathcal{M}, m/h \models \neg\Box\mathsf{X}\varphi$ by the def. of $[I\, dxstit]\varphi$. Hence, (3) $\mathcal{M}, m/h \models \Box(do(\alpha_I) \rightarrow \mathsf{X}\varphi)$ by the def. of $[\alpha_I]\varphi$. Since $\mathcal{M}, m/h \models \Box(\underline{do}(\alpha_I) \rightarrow do(\alpha_I))$ by the def. of $\underline{do}(\alpha_I)$, it follows that $\mathcal{M}, m/h \models \Box(\underline{do}(\alpha_I) \rightarrow \mathsf{X}\varphi)$. So, (4) $\mathcal{M}, m/h \models do(\alpha_I) \boxplus\!\!\rightarrow \varphi$ by the def. of $do(\alpha_I) \boxplus\!\!\rightarrow \varphi$. In addition, (5) $\mathcal{M}, m/h \models \mathsf{X}\varphi$ follows from (1) and (3). (1), (4), and (5) suffice to conclude that $\mathcal{M}, m/h \models [I\, pres]\varphi$.

2. If $\mathcal{M}, m/h \models [I\, sres]\varphi$, then there is $\alpha \in Ag\text{-}Acts$ s.t. $\mathcal{M}, m/h \models but(\alpha_I, \varphi)$ by the def. of $[I\, sres]\varphi$. Hence, $\mathcal{M}, m/h \models \mathsf{X}\varphi$ by the def. of $but(\alpha_I, \varphi)$. Given the def. of $[I\, sres]\varphi$ and $[I\, pres]\varphi$, this suffices to conclude that $\mathcal{M}, m/h \models [I\, pres]\varphi$.

3. Analogous to item 2.

4. Suppose that the antecedent is true at m/h. Since $\mathcal{M}, m/h \models [I\, dxstit]\varphi$, $\mathcal{M}, m/h \models [I\, pres]\varphi$ by item 1. Hence, it suffices to prove that $\mathcal{M}, m/h \models ness(\alpha_I, \varphi)$ where $\alpha_I = \mathbf{act}(m/h)_I$. But this is an immediate consequence of the hypothesis, since $\mathcal{M}, m/h \models [I\, dxstit]\varphi$ implies that $\mathcal{M}, m/h \models [\alpha_I]\varphi$ by the def. of $[I\, dxstit]\varphi$.

5. Suppose that the antecedent is true at m/h. Since $\mathcal{M}, m/h \models [I\, pres]\varphi$, $\mathcal{M}, m/h \models \neg\Box\mathsf{X}\varphi$ by the def. of $[I\, pres]\varphi$. Hence, it suffices to prove that $\mathcal{M}, m/h \models \Box(do(\alpha_I) \rightarrow \mathsf{X}\varphi)$ where $\alpha_I = \mathbf{act}(m/h)_I$. But this is an immediate consequence of the hypothesis, since $\mathcal{M}, m/h \models [I\, pres]\varphi$ implies that $\mathcal{M}, m/h \models \Box(\underline{do}(\alpha_I) \rightarrow \mathsf{X}\varphi)$ by the def. of $[I\, pres]\varphi$ and $do(\alpha_I) \boxplus\!\!\rightarrow \varphi$.

□

A.2.2. Proposition. *Where* $I \subseteq Ag$ *and* $\varphi \in \mathcal{L}_{\mathsf{ALO}_n}$, *the following are* invalid *in the class of* ALO_n *frames:*

1. $[I\, sres]\varphi \rightarrow [I\, dxstit]\varphi$

2. $[I\, sres]\varphi \rightarrow [I\, res]\varphi$

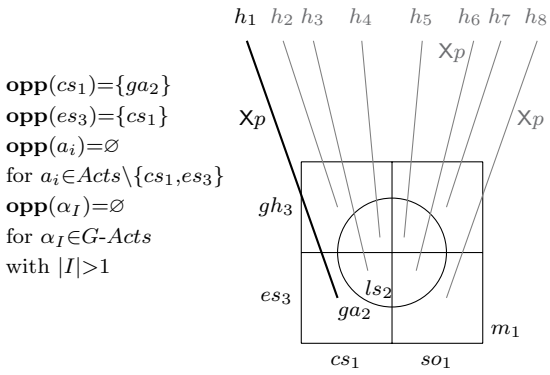


Figure A.1: An ALO_n model to prove Proposition A.2.2

3. $[I\, pres]\varphi \to [I\, dxstit]\varphi$
4. $[I\, pres]\varphi \to [I\, res]\varphi$
5. $[I\, pres]\varphi \to [I\, sres]\varphi$
6. $([I\, dxstit]\varphi \vee [I\, res]\varphi) \to [I\, sres]\varphi$

Proof:
Let $\mathcal{M}$ be the ALO_n model depicted in Figure A.1. We check that the above implications are false at some index in $\mathcal{M}$.

1. $\mathcal{M}, m_1/h_1 \models [3\, sres]p$ and $\mathcal{M}, m_1/h_1 \not\models [3\, dxstit]p$.

   To see that $\mathcal{M}, m_1/h_1 \models [3\, sres]p$, notice that agent 3 performs action es_3 at m_1/h_1 and that this action is done unopposed at m_1/h_6 and m_1/h_8. Since $\mathsf{X}p$ is true at both the latter indices, (1) $do(es_3) \boxplus\!\!\rightarrow p$ is true at m_1/h_1. In addition, the only index at which all agents behave as they do at m_1/h_1 except for agent 3 is m_1/h_2, where $\mathsf{X}p$ is false. Since $\mathsf{X}p$ is true at m_1/h_1, it follows that (2) $but(es_3, p)$ is also true at m_1/h_1. Finally, at m_1/h_3 agent 3 performs action gh_3 and $\mathsf{X}p$ is false, so (3) $\neg\Box\mathsf{X}p$ and (4) $\neg\Box do(es_3)$ are true at m_1/h_1. (1) to (4) suffice to conclude: $\mathcal{M}, m_1/h_1 \models [3\, sres]p$.

   To see that $\mathcal{M}, m_1/h_1 \not\models [3\, dxstit]p$ observe that $\mathbf{act}(m_1/h_3)(3) = \mathbf{act}(m_1/h_1)(3) = es_3$ and $\mathcal{M}, m_1/h_3 \not\models \mathsf{X}p$. This means that the action performed by agent 3 at m_1/h_3 does not guarantees the truth of $\mathsf{X}p$.

2. Let $\mathcal{M}'$ be the ALO_n model obtained from $\mathcal{M}$ by replacing the actions available to agent 2 with a vacuous action vc_2, i.e., by letting $\mathbf{act}(m_1/h)(2) = vc_2$ for all $h \in H_{m_1}$. Then, $\mathcal{M}', m_1/h_1 \models [3\, sres]p$ and $\mathcal{M}', m_1/h_1 \not\models [3\, res]p$.

The reasoning to see that $\mathcal{M}', m_1/h_1 \models [3\,sres]p$ is analogous to the reasoning in item 1. To see that $\mathcal{M}', m_1/h_1 \not\models [3\,res]p$ observe that neither es_3 (hence, es_3vc_2) nor es_3c_1 (hence, $es_3c_1vc_2$) are sufficient conditions for p because these actions are performed at m_1/h_3 where $\mathsf{X}p$ is false. This means that there is no actual condition for p that is (minimally) sufficient for p.

3. Since $[I\,sres]\varphi$ entails $[I\,pres]\varphi$ by Prop. A.2.1(2), we can infer from item 1 that $\mathcal{M}, m_1/h_1 \models [3\,pres]p$ and $\mathcal{M}, m_1/h_1 \not\models [3\,dxstit]p$.

4. Since $[I\,sres]\varphi$ entails $[I\,pres]\varphi$ by Prop. A.2.1(2), we can infer from item 2 that $\mathcal{M}', m_1/h_1 \models [3\,pres]p$ and $\mathcal{M}', m_1/h_1 \not\models [3\,res]p$.

5. $\mathcal{M}, m_1/h_8 \models [\{1,2,3\}\,pres]p$ and $\mathcal{M}, m_1/h_8 \not\models [\{1,2,3\}\,res]p$.

   To see that $\mathcal{M}, m_1/h_8 \models [\{1,2,3\}\,pres]p$ observe that the group of all agents perform action $so_1ga_2es_3$ at m_1/h_8. This action is done unopposed at m_1/h_8 where $\mathsf{X}p$ is true. Hence, (1) $do(o_1ga_2es_3) \boxplus\!\!\rightarrow p$ is true at m_1/h_8. In addition, since $\mathsf{X}p$ is false at m_1/h_2, (2) $\neg\Box\mathsf{X}p$ is true at m_1/h_8. Given that $\mathsf{X}p$ is true at m_1/h_8, (1) and (2) suffice to conclude: $\mathcal{M}, m_1/h_8 \models [\{1,2,3\}\,pres]p$.

   To see that $\mathcal{M}, m_1/h_8 \not\models [\{1,2,3\}\,res]p$, observe that the action performed by the group $\{1,2\}$ at m_1/h_8, namely so_1es_3, is a sufficient condition for p. In fact, so_1es_3 is performed at m_1/h_6 and at m_1/h_8 and $\mathsf{X}p$ is true at both these indices. Hence, the action performed by the group of all agents at m_1/h_8 is not a *minimal* sufficient condition for p.

6. $\mathcal{M}, m_1/h_8 \models [\{1,3\}\,dxstit]p \wedge [\{1,3\}\,res]p$ and $\mathcal{M}, m_1/h_8 \not\models [\{1,3\}\,sres]p$.

   To see that $\mathcal{M}, m_1/h_8 \models [\{1,3\}\,dxstit]p \wedge [\{1,3\}\,res]p$, it suffices to see that the action so_1es_3, which is performed by the group $\{1,3\}$ at m_1/h_8, is a minimal sufficient condition for p. We have seen that this action is a sufficient condition for p in item 5. The fact that so_1es_3 is a *minimal* sufficient condition for p follows from the fact that so_1 is performed at m_1/h_7 and es_3 is performed at m_1h_3 and $\mathsf{X}p$ is false at these indices.

   To see that $\mathcal{M}, m_1/h_8 \not\models [\{1,3\}\,sres]p$ observe that cs_1es_3 is an alternative action available to group $\{1,3\}$ that would guarantee the truth of $\mathsf{X}p$ given what agent 2 does at m_1/h_8: the global action $cs_1ga_2es_3$ is performed at m_1/h_1 where $\mathsf{X}p$ is true.

□

Appendix B

Appendix of Chapter 4

In this appendix we prove the main propositions from Chapter 4.3.2, i.e., Propositions 4.3.9 and 4.3.11.

B.1 Proof of Proposition 4.3.9

Let $\mathcal{M} = \langle Mom, m_0, <, \mathbf{act}, \mathbf{dev}, \preceq, \pi \rangle$ be either a rewind model or an independence model and m/h an index in $\mathcal{M}$. The validity of $\mathsf{Dis}_{\mathsf{X}}$ and $\mathsf{Dis}_{\mathsf{Y}}$ is a direct consequence of the evaluation rule for $\Box\!\!\rightarrow$ [Def. 4.3.1] and the properties of instants [see pp. 80-81]. We only present the proof for the left-to-right direction of $\mathsf{Dis}_{\mathsf{X}}$ as an illustration. Recall that, for any $h \in H_m$, $succ_h(m)$ is the successor of m on history h and that t_m is the instant to which m belongs. The definition of $Inst$ [Def. 4.2.3] ensures that, for any $h \in H_m$ and $h' \in H_{m'}$,

1. $\mathsf{t}_m = \mathsf{t}_{m'}$ iff $\mathsf{t}_{succ_h(m)} = \mathsf{t}_{succ_{h'}(m')}$.

We will repeatedly use this fact in the proofs below without explicit mention. The proof of the validity of the left-to-right direction of $\mathsf{Dis}_{\mathsf{X}}$ is as follows:

($\mathsf{Dis}_{\mathsf{X}}$, L-R) Suppose that $\mathcal{M}, m/h \models \mathsf{X}(\varphi \,\Box\!\!\rightarrow \psi)$. By the def. of truth [Def. 4.2.6], $\mathcal{M}, succ_h(m)/h \models \varphi \,\Box\!\!\rightarrow \psi$. By the semantics of $\Box\!\!\rightarrow$ [Def. 4.3.1], there are two cases:

(i) There is no $h' \in Hist$ s.t. $\mathcal{M}, \mathsf{t}_{succ_h(m)}/h' \models \varphi$.

We want to show that there is no $h' \in Hist$ s.t. $\mathcal{M}, \mathsf{t}_m/h' \models \mathsf{X}\varphi$. Suppose, toward contradiction, that there is such h'. Then, by Def. 4.2.6, $\mathcal{M}, \mathsf{t}_{succ_h(m)}/h' \models \varphi$, against the hypothesis (i). So, there is no $h' \in Hist$ s.t. $\mathcal{M}, \mathsf{t}_m/h' \models \mathsf{X}\varphi$. By Def. 4.3.1 (i), we conclude: $\mathcal{M}, m/h \models \mathsf{X}\varphi \,\Box\!\!\rightarrow \mathsf{X}\psi$.

(ii) There is $h' \in Hist$ s.t. $\mathcal{M}, \mathsf{t}_{succ_h(m)}/h' \models \varphi \wedge \psi$ and, for all $h'' \in Hist$ s.t. $\mathcal{M}, \mathsf{t}_{succ_h(m)}/h'' \models \varphi \wedge \neg\psi$, $h'' \not\preceq_h h'$.

By Def. 4.2.6, (1) $\mathcal{M}, \mathsf{t}_m/h' \models \mathsf{X}\varphi \wedge \mathsf{X}\psi$. Take any $h^* \in Hist$ s.t. (2) $\mathcal{M}, \mathsf{t}_m/h^* \models \mathsf{X}\varphi \wedge \neg\mathsf{X}\psi$. We want to show that $h^* \not\preceq_h h'$. By Def. 4.2.6, (2) implies that $\mathcal{M}, \mathsf{t}_{succ_h(m)}/h^* \models \varphi \wedge \neg\psi$, and so (3) $h^* \not\preceq_h h'$ by hypothesis (ii). Given Def. 4.3.1 (ii), (1) and (3) suffice to conclude: $\mathcal{M}, m/h \models \mathsf{X}\varphi \,\Box\!\!\rightarrow \mathsf{X}\psi$.

The proofs that Cen1 and Cen2 are valid are as follows:

(Cen1) Assume that $\mathcal{M}, m/h \models \Diamond\varphi \wedge \Diamond\psi$. Then, (1) there is $h' \in H_m$ s.t. $\mathcal{M}, m/h' \models \varphi \wedge \Diamond\psi$. Take any $h'' \in H_m$. We want to show that $\mathcal{M}, m/h'' \models \varphi \,\Box\!\!\rightarrow \Diamond\psi$. Given (1), the vacuous case is excluded. So, consider any history h^* s.t. (2) $\mathcal{M}, \mathsf{t}_m/h^* \models \varphi \wedge \neg\Diamond\psi$. We want to show that $h^* \not\preceq_{h''} h'$. Since $\mathcal{M}, m/h'' \models \Diamond\psi$ and $\mathcal{M}, \mathsf{t}_m/h^* \not\models \Diamond\psi$, $h^* \notin H_m$, i.e., h^* branches off from h'' earlier than m. Since $h' \in H_m$, this means that $past\_ov(h'', h') \supset past\_ov(h'', h^*)$, and so (3) $h^* \not\preceq_{h''} h'$ by Def. 4.3.3. Since h^* is an arbitrary history satisfying (2), (1) and (3) suffice to conclude that $\mathcal{M}, m/h'' \models \varphi \,\Box\!\!\rightarrow \psi$, and so $\mathcal{M}, m/h \models \Box(\varphi \,\Box\!\!\rightarrow \psi)$ (as h'' is an arbitrary history in H_m).

(Cen2) Assume that $\mathcal{M}, m/h \models \Diamond\varphi \wedge (\varphi \,\Box\!\!\rightarrow \Box\psi)$. Then, (1) there is $h' \in H_m$ s.t. $\mathcal{M}, m/h' \models \varphi$. Hence, $\varphi \,\Box\!\!\rightarrow \Box\psi$ is not vacuously true at m/h: (2) there is $h'' \in Hist$ s.t. $\mathcal{M}, \mathsf{t}_m/h'' \models \varphi \wedge \Box\psi$ and (3) for all $h^* \in Hist$ s.t. $\mathcal{M}, \mathsf{t}_m/h^* \models \varphi \wedge \neg\Box\psi$, $h^* \not\preceq_h h''$. We want to show that $\mathcal{M}, m/h \models \Box\psi$. Suppose, toward contradiction, that (∗) $\mathcal{M}, m/h \not\models \Box\psi$. Then, $h'' \notin H_m$, as $\mathcal{M}, \mathsf{t}_m/h'' \models \Box\psi$. That is: h'' branches off from h earlier than m. Since $h' \in H_m$, this means that $past\_ov(h, h') \supset past\_ov(h, h'')$, and so $h' \preceq_h h''$ by Def. 4.3.3. But, by (1) and (∗), $\mathcal{M}, m/h' \models \varphi \wedge \neg\Box\psi$, which entails $h' \not\preceq_h h''$ by (3). Hence, $\mathcal{M}, m/h \models \Box\psi$.

B.2 Proof of Proposition 4.3.11

We start by proving that $\mathsf{Exp}_\Box$ is valid in any rewind model:

Let $\mathcal{M} = \langle Mom, m_0, <, \mathbf{act}, \mathbf{dev}, \preceq^R \rangle$ be a rewind model and m/h any index in $\mathcal{M}$. Assume that $\mathcal{M}, m/h \models \Box\neg\varphi \wedge (\varphi \,\Box\!\!\rightarrow \Box\psi)$. There are two cases:

(i) There is no $h' \in Hist$ s.t. $\mathcal{M}, \mathsf{t}_m/h' \models \varphi$.

Then, for any $h' \in H_m$, $\varphi \,\Box\!\!\rightarrow \psi$ is vacuously true at m/h'. Hence, $\mathcal{M}, m/h \models \Box(\varphi \,\Box\!\!\rightarrow \psi)$ by Def. 4.2.6.

(ii) There is $h' \in Hist$ s.t. (1) $\mathcal{M}, \mathsf{t}_m/h' \models \varphi \wedge \Box\psi$ and (2) for all $h'' \in Hist$ s.t. $\mathcal{M}, \mathsf{t}_m/h'' \models \varphi \wedge \neg\Box\psi$, $h'' \not\preceq^R_h h'$.

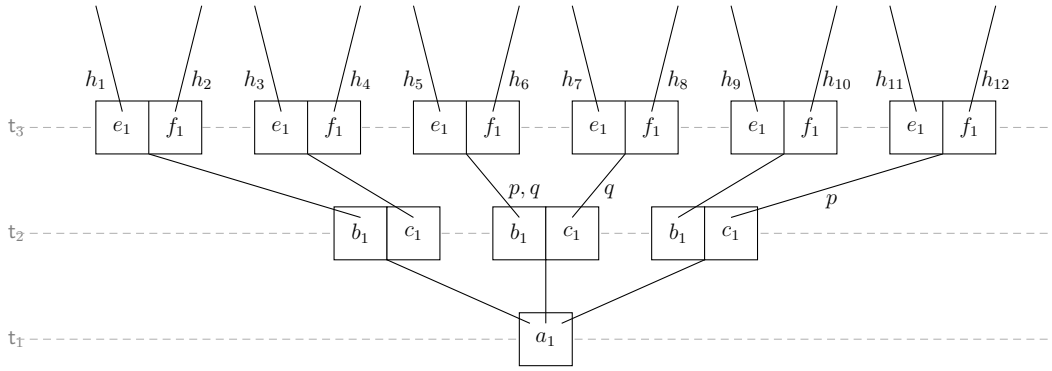


Figure B.1: An independence model not satisfying $\mathsf{Exp}_{\Box}$.

Take any $h^* \in H_m$. We want to show that $\mathcal{M}, m/h^* \models \varphi \,\Box\!\!\rightarrow \psi$. By (1), we know that (3) $\mathcal{M}, \mathsf{t}_m/h' \models \varphi \wedge \psi$. So, consider any $h'' \in Hist$ s.t. (4) $\mathcal{M}, \mathsf{t}_m/h'' \models \varphi \wedge \neg\psi$. We prove that $h'' \not\preceq^R_{h^*} h'$. Observe that:

(a) $h'' \not\preceq^R_h h'$. In fact, $\mathcal{M}, \mathsf{t}_m/h'' \models \varphi \wedge \neg\Box\psi$ by (4), and so $h'' \not\preceq^R_h h'$ by (2).

(b) Since $h, h^* \in H_m$, for any $m' < m$, h and h^* are undivided at m'.

(c) $past\_ov(h, h'') = past\_ov(h^*, h'')$. In fact, $\mathcal{M}, m/h \models \Box\neg\varphi$ by hypothesis, while $\mathcal{M}, \mathsf{t}_m/h'' \not\models \Box\neg\varphi$ (as φ is true at t_m/h''). Hence, h'' must branch off from h earlier than m. But, by (b), any history branching off from h at $m' < m$ also branches off from h^* at m'.

(d) $past\_ov(h, h') = past\_ov(h^*, h')$: analogous to (c).

(e) $num\_sep(h, h'') = num\_sep(h^*, h'')$ and $num\_sep(h, h') = num\_sep(h^*, h')$: analogous to (c).

By applying Def. 4.3.3, it is easy to see that (a), (c), (d), and (e) imply: $h'' \not\preceq^R_{h^*} h'$. Since h^* is an arbitrary history in H_m, we conclude: $\mathcal{M}, m/h \models \Box(\varphi \,\Box\!\!\rightarrow \psi)$.

We now show that there is an independence model on which $\mathsf{Exp}_{\Box}$ is invalid:

Consider Figure B.1. Assume that: (1) for any agent $i \in Ag \setminus \{1\}$ and moment m', $Acts_i^{m'} = \{vc_i\}$, (2) for any moment m' not depicted in the figure $Acts_1^{m'} = \{vc_1\}$, and (3) for any moment m', $\mathbf{dev}(m') = \varnothing$. It is not difficult to check that the defined structure is an ALD_n frame. As shown in the figure, let p be true at $\mathsf{t}_2/h_5, \mathsf{t}_2/h_6$, $\mathsf{t}_2/h_{11}, \mathsf{t}_2/h_{12}$ and q be true at t_2/h_5, t_2/h_6, t_2/h_7, and t_2/h_8. Then, $\mathsf{t}_2/h_1 \models \Box\neg p$ and $\mathsf{t}_2/h_1 \models p \,\Box\!\!\rightarrow \Box q$. In fact, (1) the most similar history to h_1 where p is true at time

$\mathtt{t}_2$ is h_5, as all unconstrained agents (i.e., all agents) do the same types of action on h_1 and h_5 at all times, and (2) $\Box q$ is true at $\mathtt{t}_2/h_5$. On the other hand, $\mathtt{t}_2/h_1 \not\models \Box(p \,\Box\!\!\rightarrow q)$. Consider, in fact, history h_3: The most similar history to h_3 where p is true at time $\mathtt{t}_2$ is h_{11}, as all unconstrained agents do the same types of action on h_3 and h_{11} at all times. Since q is false at $\mathtt{t}_2/h_{11}$, $\mathtt{t}_2/h_3 \not\models p \,\Box\!\!\rightarrow q$. Therefore, $\mathtt{t}_2/h_1 \not\models \Box(p \,\Box\!\!\rightarrow q)$.

B.2.1. Remark. The model depicted in Figure B.1 satisfies the conditions of uniformity of menus and of identity of overlapping menus [cf. items 1 and 2 on page 100]. Hence, $\mathsf{Exp}_{\Box}$ remains invalid in the class of independence models satisfying these conditions.

Bibliography

Carlos E. Alchourrón, Peter Gärdenfors, and David Makinson. On the logic of theory change: Partial meet functions for contraction and revision. *Journal of Symbolic Logic*, 50(2):510–30, 1985.

Maria Aloni. Free choice, modals, and imperatives. *Natural Language Semantics*, 15(1):65–94, 2007.

Maria Aloni. FC disjunction in state-based semantics. Institute for Logic, Language and Computation, University of Amsterdam, 2018.

Maria Aloni and Ivano Ciardelli. A logical account of free choice imperatives. In *The Dynamic, Inquisitive, and Visionary life of φ, $?\varphi$, and $\diamond\varphi$*, pages 1–17. Institute for Logic, Language and Computation, 2013.

Rajeev Alur, Thomas A. Henzinger, and Orna Kupferman. Alternating-time temporal logic. *Journal of the ACM*, 49(5):672–713, 2002.

Alan Ross Anderson. A Reduction of Deontic Logic to Alethic Modal Logic. *Mind*, LXVII(265):100–103, 1958.

Albert J.J. Anglberger. Dynamic deontic logic and its paradoxes. *Studia Logica: An International Journal for Symbolic Logic*, 89(3):427–435, 2008.

Albert J.J. Anglberger, Federico L.G. Faroldi, and Johannes Korbmacher. An exact truthmaker semantics for obligation and permission. In *Deontic Logic and Normative Sytems, 13th International Conference (DEON 2016)*, pages 16–31. College Publications, Milton Keynes, 2016.

Sergei Artemov. The logic of justification. *The Review of Symbolic Logic*, 1(4): 477–513, 2008.

Robert J. Aumann. Backward induction and common knowledge of rationality. *Games and Economic Behavior*, 8(1):6–19, 1995.

Philippe Balbiani, Olivier Gasquet, Nicolas Troquard, and François Schwarzentruber. Coalition games over kripke semantics: Expressivity and complexity. In Cédric Dégremont, Laurent Keiff, and Helge Rückert, editors, *Dialogues, Logics and Other Strange Things: Essays in Honour of Shahid Rahman*, Tributes, pages 11–32. College Publications, 2008a.

Philippe Balbiani, Andreas Herzig, and Nicolas Troquard. Alternative axiomatics and complexity of deliberative stit theories. *Journal of Philosophical Logic*, 37 (4):387–406, 2008b.

Alexandru Baltag, Lawrence S. Moss, and Slawomir Solecki. The logic of public announcements, common knowledge, and private suspicious. In Itzhak Gilboa, editor, *Proceedings of the 7th Conference on Theoretical Aspects of Rationality and Knowledge (TARK 1998)*, pages 43–56. Morgan Kaufmann, Evanston, IL, 1998.

Alexandru Baltag, Bryan Renne, and Sonja Smets. The logic of justified belief change, soft evidence and defeasible knowledge. In Luke Ong and Ruy de Queiroz, editors, *Logic, Language, Information and Computation*, pages 168–190. Springer Berlin Heidelberg, Berlin, Heidelberg, 2012.

Alexandru Baltag, Bryan Renne, and Sonja Smets. The logic of justified belief, explicit knowledge, and conclusive evidence. *Annals of Pure and Applied Logic*, 165(1):49–81, 2014.

Alexandru Baltag, Ilaria Canavotto, and Sonja Smets. Causal agency and responsibility: A refinement of STIT logic. In Alessandro Giordani and Jacek Malinowski, editors, *Logic in High Definition, Trends in Logical Semantics*, volume 56 of *Trends in Logic*. Springer, Berlin, Forthcoming.

Pierpaolo Battigalli. On rationalizability in extensive games. *Journal of Economic Theory*, 74(1):40–61, 1997.

Pierpaolo Battigalli and Marciano Siniscalchi. Strong belief and forward induction reasoning. *Journal of Economic Theory*, 106(2):356–391, 2002.

Michael Baumgartner. A regularity theoretic approach to actual causation. *Erkenntnis*, 73:85–109, 2013.

Helen Beebee, Christopher Hitchcock, and Peter Menzies. *The Oxford Handbook of Causation*. Oxford University Press, 2010.

Nuel Belnap and Michael Perloff. Seeing to it that: A canonical form for agentives. *Theoria*, 54(3):175–199, 1988.

Nuel Belnap and Michael Perloff. In the realm of agents. *Annals of Mathematics and Artificial Intelligence*, 9(1):25–48, 1993.

Nuel Belnap, Michael Perloff, and Ming Xu. *Facing the Future: Agents and Choices in Our Indeterministic World*. Oxford University Press, Oxford, 2001.

Jonathan Bennett. *A Philosophical Guide to Conditionals*. Clarendon Press, Oxford, 2003.

Johan van Benthem. *Logical Dynamics of Information and Interaction*. Cambridge University Press, Cambridge, 2011.

Johan van Benthem and Fenrong Liu. Deontic logic and preference change. *IfColog*, 1(2):1–46, 2014.

Johan van Benthem and Eric Pacuit. Connecting logics of choice and change. In Thomas Müller, editor, *Nuel Belnap on Indeterminism and Free Action*, pages 291–314. Springer International Publishing, Cham, 2014.

Johan van Benthem and Fernando R Velázquez-Quesada. The dynamics of awareness. *Synthese*, 177(1):5–27, 2010.

Johan van Benthem, Jan van Eijck, and Barteld Kooi. Logics of communication and change. *Information and Computation*, 204(11):1620–1662, 2006.

Johan van Benthem, Davide Grossi, and Fenrong Liu. Priority structures in deontic logic. *Theoria*, 80(2):116–152, 2014.

Sara Bernstein. Causal proportions and moral responsibility. In David Shoemaker, editor, *Oxford Studies in Agency and Responsibility. Volume 4*, chapter 9, pages 165–182. Oxford University Press, 2017.

Francesco Berto. Aboutness in imagination. *Philosophical Studies*, 175:1871–86, 2018.

Cristina Bicchieri. Strategic behavior and counterfactuals. *Synthese*, 76:135–169, 1988.

Patrick Blackburn, Maarten de Rijke, and Yde Venema. *Modal Logic*. Cambridge Tracts in Theoretical Computer Science. Cambridge University Press, 2001.

Giacomo Bonanno. A doxastic behavioral characterization of generalized backward induction. *Games and Economic Behavior*, 88:221–241, 2014.

Giacomo Bonanno. Reasoning about strategies and rational play in dynamic games. In Johan van Banthem, Sujata Ghosh, and Rineke Verbrugge, editors, *Models of Strategic Reasoning. Logics, Games, and Communities*, pages 34–62. Springer-Verlag, Berlin, Heidelberg, 2015.

Matthew Braham and Martin van Hees. Responsibility voids. *The Philosophical Quarterly*, 61(242):6–15, 2011.

Matthew Braham and Martin van Hees. An anatomy of moral responsibility. *Mind*, 121(483):601–634, 2012.

Matthew Braham and Martin van Hees. Voids or fragmentation: Moral responsibility for collective outcomes. *The Economic Journal*, 128(612):F95–F113, 2018.

Jan M. Broersen. *Modal Action Logics for Reasoning about Reactive Systems*. PhD thesis, Vrije Universiteit Amsterdam, 2003.

Jan M. Broersen. Action negation and alternative reductions for dynamic deontic logics. *Journal of Applied Logic*, 2(1 SPEC. ISS.):153–168, 2004.

Jan M. Broersen. A complete STIT logic for knowledge and action, and some of its applications. In Matteo Baldoni, Tran Cao Son, Birna M. van Riemsdijk, and Michael Winikoff, editors, *Declarative Agent Languages and Technologies VI*, pages 47–59. Springer Berlin Heidelberg, Berlin, Heidelberg, 2009.

Jan M. Broersen. Making a start with the *stit* logic analysis of intentional action. *Journal of Philosophical Logic*, 40(4):499–530, 2011a.

Jan M. Broersen. Deontic epistemic *stit* logic distinguishing modes of mens rea. *Journal of Applied Logic*, 9(2):137–152, 2011b.

Jan M. Broersen. Probabilistic stit logic and its decomposition. *International Journal of Approximate Reasoning*, 54:467–477, 2013.

Jan M. Broersen. A *stit* logic analysis of morally lucky and legally lucky action outcomes. In Thomas Müller, editor, *Nuel Belnap on Indeterminism and Free Action*, pages 75–98. Springer International Publishing, Cham, 2014a.

Jan M. Broersen. On the reconciliation of logics of agency and logics of event types. *Outstanding Contributions to Logic*, 1(Krister Segerberg on Logic of Actions):41–59, 2014b.

Jan M. Broersen and Andreas Herzig. Using STIT theory to talk about strategies. In Johan van Benthem, Sujata Ghosh, and Rineke Verbrugge, editors, *Models of Strategic Reasoning. Logics, Games, and Communities*, pages 137–173. Springer-Verlag, Berlin, Heidelberg, 2015.

Jan M. Broersen and Aldo I. Ramírez Abarca. Knowledge and subjective oughts in STIT logic. In Jan M. Broersen, Cleo Condoravdi, Shyam Nair, and Gabriella Pigozzi, editors, *Deontic Logic and Normative Systems, 14th International Conference (DEON 2018)*, pages 51–69. College Publications, 2018.

Jan M. Broersen, Andreas Herzig, and Nicolas Troquard. Embedding alternating-time temporal logic in strategic STIT logic of agency. *Journal of Logic and Computation*, 16(5):559–578, 2006a.

Jan M. Broersen, Andreas Herzig, and Nicolas Troquard. From coalition logic to STIT. *Electronic Notes in Theoretical Computer Science*, 157(4):23–35, 2006b.

Mark A. Brown. On the logic of ability. *Journal of Philosophical Logic*, 17(1): 1–26, 1988.

Kimberley Brownlee. *Conscience and Conviction: The Case for Civil Disobedience*. Oxford University Press, Oxford, 2012.

Nils Bulling and Mehdi Dastani. Coalitional responsibility in strategic settings. In *Computational Logic in Multi-Agent Systems. CLIMA 2013*, volume 8143 of *Lecture Notes in Computer Science*, pages 172–189. Springer-Verlag, Berlin, Heidelberg, 2013.

Ruth M.J. Byrne. *The Rational Imagination. How People Create Alternatives to Reality*. MIT Press, Cambridge, Mass., 2005.

Ruth M.J. Byrne and Vittorio Girotto. Cognitive processes in counterfactual thinking. In Keith D. Markman, William M.P. Klein, and Julie A. Suhr, editors, *Handbook of Imagination and Mental Simulation*, pages 151–60. Taylor and Francis, New York, 2009.

Ilaria Canavotto and Alessandro Giordani. Normative conflicts in a dynamic logic of norms and codes. In Jan M. Broersen, Cleo Condoravdi, Shyam Nair, and Gabriella Pigozzi, editors, *Deontic Logic and Normative Systems, 14th International Conference (DEON 2018)*, pages 71–90. College Publications, Milton Keynes, 2018.

Ilaria Canavotto and Alessandro Giordani. Enriching deontic logic. *Journal of Logic and Computation*, pages 241–263, 2019.

Ilaria Canavotto and Eric Pacuit. Choice-driven counterfactuals. Manuscript in preparation. Institute for Logic, Language and Computation, University of Amsterdam and Department of Philosophy, University of Maryland, 2020.

Ilaria Canavotto, Francesco Berto, and Alessandro Giordani. Voluntary imagination: A fine-grained analysis. *The Review of Symbolic Logic*, pages 1–34, 2020.

José Carmo and Andrew J.I. Jones. Deontic logic and different levels of ideality. RRDMIST 1/95, 1995.

José Carmo and Andrew J.I. Jones. A new approach to contrary-to-duty obligations. In Donald Nute, editor, *Defeasible Deontic Logic*, pages 317–344. Springer Netherlands, Dordrecht, 1997.

José Carmo and Andrew J.I. Jones. Deontic logic and contrary-to-duties. In Dov Gabbay and Franz Guenthner, editors, *Handbook of Philosophical Logic: Volume 8*, chapter 4, pages 265–343. Springer Netherlands, Dordrecht, 2002.

Claudia Carr and Maureen Johnson. *Beginning Criminal Law*. Routledge, London and New York, 2013.

David J. Chalmers. Does conceivability entail possibility? In Tamar S. Gendler and John Hawthorne, editors, *Conceivability and Possibility*, pages 145–99. Oxford University Press, Oxford, 2002.

Brian F. Chellas. *The Logical Form of Imperatives*. PhD thesis, Stanford University, 1969.

Brian F. Chellas. Time and modality in the logic of agency. *Studia Logica*, 51(3): 485–517, 1992.

Roderick M. Chisholm. Contrary-to-duty imperatives and deontic logic. *Analysis*, 24(2):33–36, 1963.

Roberto Ciuni and John F. Horty. Stit logics, games, knowledge, and freedom. *Outstanding Contributions to Logic*, 5(Johan van Benthem on Logic and Information Dynamics):631–656, 2014.

Roberto Ciuni and Emiliano Lorini. Comparing semantics for temporal STIT logic. *Logique et Analyse*, 61(243):299–339, 2018.

Roberto Ciuni and Rosja Mastop. Attributing distributed responsibility in Stit logic. In Xiangdong He, John F. Horty, and Eric Pacuit, editors, *Logic, Rationality, and Interaction*, pages 66–75. Springer Berlin Heidelberg, Berlin, Heidelberg, 2009.

Philip R. Cohen and Hector J. Levesque. Intention is choice with commitment. *Artificial Intelligence*, 42(2):213–261, 1990.

Robert Craven and Marek Sergot. Agent strands in the action language $n\mathcal{C}+$. *Journal of Applied Logic*, 6(2):172–191, 2008.

Ron van der Meyden. The dynamic logic of permission. *Journal of Logic and Computation*, 6(3):465–479, 1996.

Franz Dietrich and Christian List. Reason-based choice and context-dependence: An explanatory framework. *Economics and Philosophy*, 32(2):175–229, 2016.

Frank Dignum and John-Jules Ch. Meyer. Negations of transactions and their use in the specification of dynamic and deontic integrity constraints. In Marta Zofia Kwiatkowska, Michael William Thomas, and Richard Monro Shields, editors, *Semantics for Concurrency*, pages 61–80. Springer London, London, 1990.

Hans P. van Ditmarsch, Wiebe van der Hoek, and Barteld P. Kooi. *Dynamic Epistemic Logic*. Springer, Dordrecht, 2008.

Hein Duijf. *Let's Do It! Collective Responsibility, Joint Action, and Participation.* PhD thesis, Universiteit Utrecht, 2018.

Jonathan Evans and David Over. *If*. Oxford University Press, Oxford, 2004.

Ronald Fagin, Joseph Y. Halpern, Yoram Moses, and Moshe Vardi. *Reasoning About Knowledge*. The MIT Press, Cambridge, MA, 1995.

David Fernández–Duque, Ángel Nepomuceno–Fernández, Enrique Sarrión–Morrillo, Fernando Soler–Toscano, and Fernando R. Velázquez–Quesada. Forgetting complex propositions. *Logic Journal of the IGPL*, 23(6):942–965, 2015.

Kit Fine. Angellic content. *Journal of Philosophical Logic*, 45:199–226, 2016.

Kit Fine. Truthmaker semantics. In *A Companion to the Philosophy of Language*, chapter 22, pages 556–577. John Wiley & Sons, Ltd, 2017.

Kit Fine. Compliance and command I – categorical imperatives. *The Review of Symbolic Logic*, 11(4):609–633, 2018a.

Kit Fine. Compliance and command II, imperatives and deontics. *The Review of Symbolic Logic*, 11(4):634–664, 2018b.

Michael J. Fischer and Richard E. Ladner. Propositional dynamic logic of regular programs. *Journal of Computer and System Sciences*, 18(2):194–211, 1979.

Melvin Fitting. The logic of proofs, semantically. *Annals of Pure and Applied Logic*, 132(1):1–25, 2005.

Tanya S. Fraude-Koivisto, Daniela Wuerz, and Peter M. Gollwitzer. Implementation intentions: The mental representations and cognitive procedures of if-then planning. In Keith D. Markman, William M.P. Klein, and Julie A. Suhr, editors, *Handbook of Imagination and Mental Simulation*, pages 69–87. Taylor and Francis, New York, 2009.

Tamar S. Gendler. The puzzle of imaginative resistance. *Journal of Philosophy*, 97:55–81, 2000.

Alessandro Giordani. Ability and responsibility in general action logic. In Jan M. Broersen, Cleo Condoravdi, Nair Shyam, and Gabriella Pigozzi, editors, *Deontic Logic and Normative Systems, 14th International Conference (DEON 2018)*, pages 121–138. College Publications, Milton Keynes, 2018.

Alessandro Giordani. Axiomatizing the logic of imagination. *Studia Logica*, 107: 639–57, 2019.

Alessandro Giordani and Ilaria Canavotto. Basic action deontic logic. In Olivier Roy, Allard Tamminga, and Willerd Malte, editors, *Deontic Logic and Normative Systems, 13th International Conference (DEON 2016)*, pages 80–92. College Publications, Milton Keynes, 2016.

Lou Goble. A logic for deontic dilemmas. *Journal of Applied Logic*, 3(3-4):461–483, 2005.

Lou Goble. Normative conflicts and the logic of 'ought'. *Noûs*, 43(3):450–489, 2009.

Lou Goble. Prima facie norms, normative conflicts, and dilemmas. In Dov Gabbay, John F. Horty, Xavier Parent, Ron van der Meyden, and Leendert van der Torre, editors, *Handbook of Deontic Logic and Normative Systems*, chapter 4, pages 241–352. College Publications, Milton Keynes, 2013.

Valentin Goranko. Coalition games and alternating temporal logics. In *Proceedings of the 8th Conference on Theoretical Aspects of Rationality and Knowledge*, TARK '01, pages 259–272. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2001.

Valentin Goranko and Wojciech Jamroga. Comparing semantics of logics for multi-agent systems. In *Information, Interaction and Agency*, pages 77–116. Springer Netherlands, Dordrecht, 2005.

Valentin Goranko and Solomon Passy. Using the universal modality: Gains and questions. *Journal of Logic and Computation*, 2(1):5–30, 1992.

Valentin Goranko and Govert van Drimmelen. Complete axiomatization and decidability of alternating-time temporal logic. *Theoretical Computer Science*, 353(1):93–117, 2006.

Guido Governatori and Antonino Rotolo. Logic of violations: A gentzen system for reasoning with contrary-to-duty obligations. *Australasian Journal of Logic*, 4:193–215, 2006.

Guido Governatori and Antonino Rotolo. Changing legal systems: Legal abrogations and annulments in defeasible logic. *Logic Journal of the IGPL*, 18(1): 157–194, 2010.

Guido Governatori, Francesco Olivieri, Erica Calardo, and Antonino Rotolo. Sequence semantics for norms and obligations. In Olivier Roy, Allard Tamminga, and Malte Willer, editors, *Deontic Logic and Normative Systems, 13th International Conference (DEON 2016)*, pages 93–108. College Publications, Milton Keynes, 2016.

Davide Grossi, Frank Dignum, Lambèr Royakkers, and John-Jules Ch. Meyer. Collective obligations and agents: Who gets the blame? In Alessio Lomuscio and Donald Nute, editors, *Deontic Logic in Computer Science*, pages 129–145. Springer Berlin Heidelberg, Berlin, Heidelberg, 2004.

Davide Grossi, Frank Dignum, Mehdi Dastani, and Lambèr Royakkers. Foundations of organizational structures in multiagent systems. In *Proceedings of the Fourth International Joint Conference on Autonomous Agents and Multiagent Systems*, AAMAS '05, pages 690–697. Association for Computing Machinery, New York, NY, USA, 2005.

Davide Grossi, Lambèr Royakkers, and Frank Dignum. Organizational structure and responsibility. An analysis in a dynamic logic of organizational collective agency. *Artificial Intelligence and Law*, 15(3):223–249, 2007.

Joseph Y. Halpern. Substantive rationality and backward induction. *Games and Economic Behavior*, 37(2):425–435, 2001.

Joseph Y. Halpern. Causality, responsibility, and blame: A structural-model approach. *Law, Probability and Risk*, 14:91–118, 2015.

Joseph Y. Halpern. *Actual Causality*. The MIT Press, Cambridge, MA, 2016.

Joseph Y. Halpern and Christopher Hitchcock. Actual causation and the art of modeling. In *Causality, Probability, and Heuristics: A Tribute to Judea Pearl*, chapter 23, pages 383–406. College Publications, London, 2010.

Joseph Y. Halpern and Judea Pearl. Causes and explanations: A structural-model approach. Part II: Explanations. *The British Journal for the Philosophy of Science*, 56(4):889–911, 2005.

Jörg Hansen. Deontic logics for prioritized imperatives. *Artificial Intelligence and Law*, 14(1):1–34, 2006.

Jörg Hansen. Reasoning about permission and obligation. In Sven Ove Hansson, editor, *David Makinson on Classical Methods for Non-Classical Problems*, pages 287–333. Springer Netherlands, Dordrecht, 2014.

Bengt Hansson. An analysis of some deontic logics. *Noûs*, 3:373–398, 1969.

David Harel. Dynamic logic. In Dov Gabbay and Franz Guenthner, editors, *Handbook of Philosophical Logic: Volume II: Extensions of Classical Logic*, pages 497–604. Springer Netherlands, Dordrecht, 1984.

David Harel, Dexter Kozen, and Jerzy Tiuryn. *Dynamic Logic*. The MIT Press, Cambridge, MA, 2000.

Herbert L. A. Hart and Tony Honoré. *Causation in the Law*. Oxford University Press, Oxford, 1959.

Jonathan Herring. *Criminal Law: Text, Cases, and Material*. Oxford University Press, 2012.

Adreas Herzig and Nicolas Troquard. Knowing how to play: Uniform choices in logics of agency. In *Proceedings of the 5th International Joint Conference on Autonomous Agents and Multi-agent Systems (AAMAS-06)*, pages 209–216. The Association for Computing Machinery Press, 2006.

Andreas Herzig. Logics of knowledge and action: Critical analysis and challenges. *Autonomous Agents and Multi-Agent Systems*, 29(5):719–753, 2015.

Andreas Herzig and Dominique Longin. C&L intention revisited. In *Proceedings of the Ninth International Conference on Principles of Knowledge Representation and Reasoning*, KR'04, pages 527–535. AAAI Press, 2004.

Andreas Herzig and Emiliano Lorini. A dynamic logic of agency I: STIT, capabilities and powers. *Journal of Logic, Language and Information*, 19(1):89–121, 2010.

Andreas Herzig and François Schwarzentruber. Properties of logics of individual and group agency. In Carlos Arecers and Robert Goldblatt, editors, *Advances in Modal Logic*, volume 7, pages 133–149. College Publications, 2008.

Andreas Herzig, Emiliano Lorini, and Nicolas Troquard. Action theories. In Sven Ove Hansson and Vincent F. Hendricks, editors, *Introduction to Formal Philosophy*, pages 591–607. Springer, 2018.

Risto Hilpinen, editor. *Deontic Logic: Introductory and Systematic Readings*. Springer Netherlands, Dordrecht, 1971.

Risto Hilpinen and Paul McNamara. Deontic logic: A historical survery and introduction. In *Handbook of Deontic Logic and Normative Systems*, chapter 1, pages 3–136. College Publications, 2013.

Wiebe van der Hoek and Michael Wooldridge. Logics for multiagent systems. *AI Magazine*, 33(3):92, 2012.

Antony Honoré and John Gardner. Causation in the law. In Edward N. Zalta, editor, *Stanford Encyclopedia of Philosophy*. Fall 2019 edition, 2010.

John F. Horty. An alternative stit operator. Department of Philosophy, University of Maryland, 1989.

John F. Horty. *Agency and Deontic Logic*. Oxford University Press, Oxford and New York, 2001.

John F. Horty. Reasoning with moral conflicts. *Noûs*, 37(4):557–605, 2003.

John F. Horty. Defaults with priorities. *Journal of Philosophical Logic*, 36(4): 367–413, 2007.

John F. Horty. *Reasons as Defaults*. Oxford University Press, 2012.

John F. Horty. Epistemic oughts in stit semantics. *Ergo*, 6(4):71–120, 2019.

John F. Horty and Nuel Belnap. The deliberative stit: A study of action, omission, ability, and obligation. *Journal of Philosophical Logic*, 24(6):583–644, 1995.

John F. Horty and Eric Pacuit. Action types in stit semantics. *The Review of Symbolic Logic*, 10(4):617–637, 2017.

David Hume. *An Enquiry Concerning Human Understanding*. 1748. Reprinted by Open Court Press, LaSalle, IL (1958).

Andrew J.I. Jones and Ingmar Pörn. Ideality, sub-ideality and deontic logic. *Synthese*, 65(2):275–290, 1985.

Fengkui Ju and Jan van Eijck. A temporal dynamic deontic logic. *Journal of Logic and Computation*, 29(2):265–284, 2019.

Daniel Kahneman and Amos Tversky. Choices, values, and frames. *American Psychologist*, 39:341–50, 1984.

Alex Kaiserman. 'More of a cause': Recent work on degrees of causation and responsibility. *Philosophy Compass*, 13(7):e12498, 2018.

Gil Kalai, Ariel Rubinstein, and Rani Spiegler. Rationalizing choice functions by mutltiple rationales. *Econometrica*, 70(6):2481–2488, 2002.

Stig Kanger. New foundations for ethical theory. Stockholm. Reprinted in Hilpinen, 1971, pages 36–88, 1957.

Anthony Kenny. *Will, Freedom, and Power*. Basil Blackwell, 1975.

Anthony Kenny. Human abilities and dynamic modalities. In Juha Manninen and Raimo Tuomela, editors, *Essays on Explanation and Understanding: Studies in the Foundations of Humanities and Social Sciences*, pages 209–232. Springer Netherlands, Dordrecht, 1976.

Amy Kind. Imagining under constraints. In Amy Kind and Peter Kung, editors, *Knowledge Through Imagination*, pages 145–59. Oxford University Press, Oxford, 2016.

Amy Kind and Peter Kung, editors. *Knowledge through Imagination*. Oxford University Press, Oxford, 2016.

Barteld P. Kooi and Allard Tamminga. Moral conflicts between groups of agents. *Journal of Philosophical Logic*, 37(1):1–21, 2008.

Stephen Kosslyn and Samuel T. Moulton. Mental imagery and implicit memory. In Keith D. Markman, William M.P. Klein, and Julie A. Suhr, editors, *Handbook of Imagination and Mental Simulation*, pages 35–52. Taylor and Francis, New York, 2009.

Piotr Kulicki and Robert Trypuz. Connecting actions and states in deontic logic. *Studia Logica*, 105:915–942, 2017.

David A. Lagnado and Tobias Gerstenberg. Causation in legal and moral reasoning. In Michael R. Waldmann, editor, *The Oxford Handbook of Causal Reasoning*, chapter 29, pages 565–601. Oxford University Press, 2017.

Peter Langland-Hassan. On choosing what to imagine. In Amy Kind and Peter Kung, editors, *Knowledge Through Imagination*, pages 61–84. Oxford University Press, Oxford, 2016.

David Lewis. *Counterfactuals*. Blackwell, Oxford, 1973a.

David Lewis. Causation. *The Journal of Philosophy*, 70(17):556–567, 1973b.

David Lewis. Semantic analyses for dyadic deontic logic. In Sören Stenlund, Ann-Mari Henschen-Dahlquist, Lars Lindahl, Lennart Nordenfelt, and Jan Odelstad, editors, *Logical Theory and Semantic Analysis: Essays Dedicated to Stig Kanger on His Fiftieth Birthday*, pages 1–14. Springer Netherlands, Dordrecht, 1974.

David Lewis. Counterfactual dependence and time's arrow. *Noûs*, 13(4):455–476, 1979.

David Lewis. Causal explanation. In *Philosophical Papers, Volume II*, pages 214–240. Oxford University Press, Oxford, 1986.

David Lewis. Relevant implication. *Theoria*, 54(3):161–174, 1988.

Fenrong Liu. *Reasoning about Preference Dynamics*. Number 354 of the Synthese Library. Springer, Amsterdam, 2011.

Emiliano Lorini. Temporal STIT logic and its application to normative reasoning. *Journal of Applied Non-Classical Logics*, 23(4):372–399, 2013.

Emiliano Lorini and Andreas Herzig. A logic of intention and attempt. *Synthese*, 163(1):45–77, 2008.

Emiliano Lorini and Giovanni Sartor. A STIT logic for reasoning about social influence. *Studia Logica*, 104(4):773–812, aug 2016.

Emiliano Lorini and François Schwarzentruber. A logic for reasoning about counterfactual emotions. *Artificial Intelligence*, 175(3):814–847, 2011.

Emiliano Lorini, Dominique Longin, and Eunate Mayor. A logical analysis of responsibility attribution: Emotions, individuals and collectives. *Journal of Logic and Computation*, 24(6):1313–1339, 2014.

John L. Mackie. Causes and conditions. *American Philosophical Quarterly*, 2(4): 245–264, 1965.

John L. Mackie. *The Cement of the Universe: A Study of Causation*. Oxford University Press, Oxford, 1974.

David Makinson and Leendert van der Torre. Input/output logics. *Journal of Philosophical Logic*, 29(4):383–408, 2000.

David Makinson and Leendert van der Torre. Constraints for input/output logics. *Journal of Philosophical Logic*, 30(2):155–185, 2001.

Maarten J. Marx. Complexity of products of modal logics. *Journal of Logic and Computation*, 9(2):197–214, 1999.

Rosja Mastop. Characterising responsibility in organisational structures: The problem of many hands. In Guido Governatori and Giovanni Sartor, editors, *Deontic Logic in Computer Science*, pages 274–287. Springer Berlin Heidelberg, Berlin, Heidelberg, 2010.

L. Thorne McCarty. Permissions and obligations. In *Proceedings of the Eighth International Joint Conference on Artificial Intelligence - Volume 1*, IJCAI'83, pages 287–294. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1983.

John-Jules Ch. Meyer. A different approach to deontic logic: Deontic logic viewed as a variant of dynamic logic. *Notre Dame Journal of Formal Logic*, 29(1):109–136, 1988.

John-Jules Ch. Meyer, Wiebe van der Hoek, and Bernardus van Linder. A logical approach to the dynamics of commitments. *Artificial Intelligence*, 113(1):1–40, 1999.

Michael S. Moore. *Causation and Responsibility: An Essary in Law, Morals, and Metaphysics*. Oxford University Press, 2009.

Michael S. Moore. Causation in the law. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Winter 201 edition, 2019.

Thomas Müller. On the formal structure of continuous action. In Renate Schmidt, Ian Pratt-Hartmann, Mark Reynolds, and Heinrich Wansing, editors, *Advances in Modal Logic, Volume 5*, pages 191–209. King's College Publications, 2005.

Stephen Mumford and Rani Lill Anjum. *Causation: A Very Short Introduction*. Oxford University Press, 2013.

Shaun Nichols and Stephen P. Stich. *Mindreading. An Integrated Account of Pretence, Self-Awareness, and Understanding Other Minds*. Oxford University Press, Oxford, 2003.

Mike Oaksford and Nick Chater, editors. *Cognition and Conditionals*. Oxford University Press, Oxford, 2010.

Martin J. Osborne and Ariel Rubinstein. *A Course in Game Theory*. MIT Press, Cambridge, MA, 1994.

Eric Pacuit and Olivier Roy. Epistemic foundations of games. In Edward N. Zalta, editor, *Stanford Encyclopedia of Philosophy*, 2015.

Xavier Parent and Leendert van der Torre. Input/output logic. In Dov Gabbay, John F. Horty, Ron van der Meyden, and Leendert van der Torre, editors, *Handbook of Deontic Logic and Normative Systems*, chapter 8, pages 241–352. College Publications, Milton Keynes, 2013.

Rohit Parikh. The completeness of propositional dynamic logic. In Józef Winkowski, editor, *Mathematical Foundations of Computer Science 1978*, volume 64 of *Lecture Notes in Computer Science*, pages 403–415. Springer Berlin Heidelberg, Berlin, Heidelberg, 1978.

Marc Pauly. *Logic for Social Software*. PhD thesis, University of Amsterdam, 2001.

Marc Pauly. A modal logic for coalitional power in games. *Journal of Logic and Computation*, 12(1):149–166, 2002.

Judea Pearl. *Causality. Models, Reasoning, and Inference.* Cambridge University Press, Cambridge, 2000.

Andres Perea. *Epistemic Game Theory: Reasoning and Choice.* Cambridge UP, 2012.

Andrés Perea. Belief in the opponents' future rationality. *Games and Economic Behavior*, 83:231 – 254, 2014.

John Perry. Possible worlds and subject matter. In *The Problem on the Essential Indexical and Other Essays*, pages 145–160. CSLI Publications, 1989.

Tomasz Placek and Thomas Müller. Counterfactuals and historical possibility. *Synthese*, 154(2):173–197, 2007.

Ingmar Pörn. *Action Theory and Social Science: Some Formal Models.* Number 120 of the Synthese Library. Reidel, Dordrecht, 1977.

Douglas W. Portmore. *Commonsense Consequentialism: Wherein Morality Meets Rationality.* Oxford University Press, Oxford, 2011.

Henry Prakken and Marek Sergot. Contrary-to-duty obligations. *Studia Logica*, 57(1):91–115, 1996.

Henry Prakken and Marek Sergot. Dyadic Deontic Logic and Contrary-to-Duty Obligations. In Donald Nute, editor, *Defeasible Deontic Logic*, pages 223–262. Springer Netherlands, Dordrecht, 1997.

Vaughan R. Pratt. Semantical considerations on Floyo-hoare Logic. In *Proceedings of the 17th IEEE Symposium on Foundations of Computer Science*, pages 109–121, Los Alamitos, CA, 1976. IEEE Computer Society.

Vaughan R. Pratt. Process Logic. In Alfred V. Aho, Stephen N. Zilles, and Barry K. Rosen, editors, *Proceedings of the 6th ACM Symposium on Principles of Programming Languages*, pages 93–100. ACM Press, New York, 1979.

Arthur Prior. *Past, Present, and Future.* Oxford University Press, 1967.

Willard V.O. Quine. *Word and Object.* MIT Press, MA, 1960.

Alf Ross. Imperatives and logic. *Philosophy of Science*, 11(1):30–46, 1944.

Lambèr Royakkers and Jesse Hughes. Blame it on me. *Journal of Philosophical Logic*, 49:315–349, 2020.

Carolina Sartorio. A new form of moral luck? In Andrei Buckareff, Carlos Moya, and Sergi Rosell, editors, *Agency, Freedom, and Moral Responsibility*, pages 134–149. Palgrave Macmillan UK, London, 2015.

François Schwarzentruber. Complexity results of stit fragments. *Studia Logica*, 100(5):1001–1045, 2012.

Krister Segerberg. A completeness theorem in the modal logic of programs. *Notices of the American Mathematical Society*, 24:A–552, 1977.

Krister Segerberg. A deontic logic of action. *Studia Logica*, 41(2):269–282, 1982.

Krister Segerberg. Getting started: Beginnings in the logic of action. *Studia Logica*, 51(3):347–378, 1992.

Krister Segerberg. Outline of a logic of action. In Frank Wolter, Heinrich Wansing, Maarten de Rijke, and Michael Zakharyaschev, editors, *Advances in Modal Logic*, volume 3, pages 365–387. World Scientific, 2002.

Reinhard Selten and Ulrike Leopold. Subjunctive conditionals in decision and game theory. In *Philosophy of Economics*, pages 191–200. Springer, 1982.

Amartya Sen. Maximization and the act of choice. *Econometrica*, 65(4):745–779, 1997.

Marek Sergot. The logic of unwitting collective agency. Department of Computing, Imperial College London, 2008.

Marek Sergot and Robert Craven. The deontic component of action language. In Lou Goble and John-Jules Ch. Meyer, editors, *Deontic Logic and Artificial Normative Systems*, pages 222–237. Springer Berlin Heidelberg, Berlin, Heidelberg, 2006.

Hyun Song Shin. Counterfactuals and a theory of equilibrium in games. In Cristina Bicchieri and Maria Luisa Dalla Chiara, editors, *Knowledge, Belief, and Strategic Interaction*, pages 397–413, 1992.

Yoav Shoham. Time for action: On the relation between time, knowledge and action. In *Proceedings of the 11th international joint conference on Artificial Intelligence, Volume 2*, pages 954–959, 1989.

Brian Skyrms. Bayesian subjunctive conditionals for games and decisions. In *Game Theory, Experience, Rationality*, pages 161–172, 1998.

Michael A. Slote. Time in counterfactuals. *The Philosophical Review*, 87(1):3–27, 1978.

William Smith. *Civil Disobedience and Deliberative Democracy*. Routledge, 2013.

Robert C. Stalnaker. A theory of conditionals. In Rescher Nicholas, editor, *Studies in Logical Theory*, pages 98–112. Basil Blackwell, Oxford, 1968.

Robert C. Stalnaker. Knowledge, belief and counterfactual reasoning in games. *Economics and Philosophy*, 12(2):133–163, 1996.

Robert C. Stalnaker. Belief revision in games: Forward and backward induction. *Mathematical Social Sciences*, 36(1):31–56, 1998.

Keith E. Stanovich and Richard F. West. Individual differences in reasoning: Implications for the rationality debate? *Behavioral and Brain Sciences*, 23(5): 645–665, 2000.

Allard Tamminga. Deontic logic for strategic games. *Erkenntnis*, 78(1):183–200, 2013.

Richmond H. Thomason. Indeterminist time and truth-value gaps. *Theoria*, 36 (3):264–281, 1970.

Richmond H. Thomason. Combinations of tense and modality. In Dov Gabbay and Franz Guenthner, editors, *Handbook of Philosophical Logic: Volume II: Extensions of Classical Logic*, pages 135–165. Springer Netherlands, Dordrecht, 1984.

Richmond H. Thomason and Anil Gupta. A theory of conditionals in the context of branching time. In William L Harper, Robert C. Stalnaker, and Glenn Pearce, editors, *IFS: Conditionals, Belief, Decision, Chance and Time*, pages 299–322. Springer Netherlands, Dordrecht, 1981.

Dennis F. Thompson. Moral responsibility of public officials: The problem of many hands. *Americam Political Science Review*, 74:905–16, 1980.

Nicolas Troquard and Laure Vieu. Towards a logic of agency and actions with duration. In *European Conference on Artificial Intelligence 2006 (ECAI'06)*, pages 775–776. IOS Press, 2006.

Paolo Turrini. Agreements as norms. In Thomas Ågotnes, Jan M. Broersen, and Dag Elgesem, editors, *Deontic Logic in Computer Science, 11th International Conference (DEON 2011)*, pages 31–45. Springer Berlin Heidelberg, Berlin, Heidelberg, 2012.

Ibo Van De Poel. Moral responsibility. In Ibo van de Poel, Lambèr Royakkers, and Sjoerd D. Zwart, editors, *Moral Responsibility and the Problem of Many Hands*, chapter 1, pages 12–49. Routledge, New York and London, 2015.

Neil Van Leeuwen. The imaginative agent. In Amy Kind and Peter Kung, editors, *Knowledge Through Imagination*, pages 85–111. Oxford University Press, Oxford, 2016.

Fernando Raymundo Velázquez-Quesada. Inference and update. *Synthese*, 169 (2):283–300, 2009.

Franz von Kutschera. Bewirken. *Erkenntnis*, 24(3):253–281, 1986.

Georg H. von Wright. An essay in deontic logic and the general theory of norms. In *Acta Philosophica Fennica, Fasc. XXI*. North-Holland Pub. Co., Amsterdam, 1968.

Georg H. von Wright. *Explanation and Understanding*. Cornell University Press, Ithaca and London, 1971.

Heinrich Wansing. On the negation of action types: Constructive concurrent PDL. In Petr Hájek, Louis Valdés-Villanueva, and Dag Westerståhl, editors, *Proceedings of the 12th International Congress of Logic, Methodology and Philosophy of Science*, pages 207–225. College Publications, 2004.

Heinrich Wansing. Tableaux for multi-agent deliberative-stit logic. In Guido Governatori, Ian Hodkinson, and Yde Venema, editors, *Advances in Modal Logic, Volume 6*, pages 503–520. College Publications, 2006.

Heinrich Wansing. Remarks on the logic of imagination. A step towards understanding doxastic control through imagination. *Synthese*, 194:2843–2861, 2017.

Roel Wieringa and John-Jules Ch. Meyer. Actors, actions, and initiative in normative system specification. *Annals of Mathematics and Artificial Intelligence*, 7(1):289–346, 1993.

Timothy Williamson. *The Philosophy of Philosophy*. Blackwell, Oxford, 2007.

Timothy Williamson. Knowing by imagining. In Amy Kind and Peter Kung, editors, *Knowledge Through Imagination*, pages 113–23. Oxford University Press, Oxford, 2016.

Stefan Wölfl. Propositional Q-logic. *Journal of Philosophical Logic*, 31(5):387–414, 2002.

Stefan Wölfl. Qualitative action theory: A comparison of the semantics of alternating time temporal logic and the Kutschera-Belnap approach to agency. In Jóse Júlio Alferes and João Leite, editors, *Logics in Artificial Intelligence (JELIA'04)*, pages 70–81. Springer Berlin Heidelberg, Berlin, Heidelberg, 2004.

James Woodward. *Making Things Happen: A Theory of Causal Explanation*. Oxford University Press, Oxford, 2003.

Michael Wooldridge. *Reasoning About Rational Agents*. The MIT Press, Cambridge, MA, 2000.

Richard W. Wright. Causation, responsibility, risk, probability, naked statistics, and proof: Pruning the bramble bush by clarifying the concepts. *Iowa Law Review*, 73:1001–1077, 1988.

Richard W. Wright. The NESS account of natural causation: A response to criticism. In Markus Stepanians and Kahmen Benedikt, editors, *Critical Essays on "Causation and Responsibility"*, chapter 14, pages 13–66. De Gruyter, 2013.

Ming Xu. On the basic logic of STIT with a single agent. *Journal of Symbolic Logic*, 60(2):459–483, 1995.

Ming Xu. Causation in branching time (I): Transitions, events and causes. *Synthese*, 112(2):137–192, 1997.

Ming Xu. Axioms for deliberative stit. *Journal of Philosophical Logic*, 27:505–552, 1998.

Ming Xu. Combinations of stit and actions. *Journal of Logic, Language and Information*, 19(4):485–503, 2010.

Ming Xu. Actions as events. *Journal of Philosophical Logic*, 41(4):765–809, 2012.

Ming Xu. Combinations of "Stit" with "Ought" and "Know". *Journal of Philosophical Logic*, 44(6):851–877, 2015.

S. Yablo. *Aboutness*. Princeton University Press, Princeton, 2014.

Stephen Yablo. Is conceivability a guide to possibility? *Philosophy and Phenomenological Research*, 53:1–42, 1993.

Tomoyuki Yamada. Logical dynamics of some speech acts that affect obligations and preferences. *Synthese*, 165(2):295–315, 2008.

Eduardo Zambrano. Counterfactual reasoning and common knowledge of rationality in normal form games. *Topics in Theoretical Economics*, 4(8), 2004.

Alberto Zanardo. Branching-time logic with quantification over branches: The point of view of modal logic. *Journal of Symbolic Logic*, 61(1):1–39, 1996.

Samenvatting

Waar verantwoordelijkheid je brengt
Logica's van agentschap, contrafactische uitspraken en normen

Dit proefschrift bestudeert logische systemen die inzichten uit de logica van agentschap, contrafactische implicaties en normen combineren. Het doel is om instrumenten te ontwikkelen om drie algemene vragen te beantwoorden met betrekking tot een formele analyse van *causale verantwoordelijkheid* (dat wil zeggen, verantwoordelijkheid voor wat er is gebeurd, ongeacht iemands intenties of overtuigingen): Hoe kunnen we de keuzevrijheid van individuen en groepen modelleren met betrekking tot het veroorzaken van bepaalde resultaten in complexe multi-agent scenario's? Wat zijn de logische eigenschappen, en kentheoretische waarde, van contrafactische implicaties over wat er in de loop van tijd gedaan kan, of had kunnen, worden? Welke regels gelden voor normatief redeneren? De eerste vraag komt voort uit het feit dat agenten alleen verantwoordelijk zijn voor wat ze veroorzaakt hebben. De tweede ontstaat omdat causale verantwoordelijkheid doorgaans wordt bepaald door te overwegen wat er zou zijn gebeurd als de relevante agenten anders gehandeld hadden. De derde komt voort uit het feit dat agenten alleen verantwoordelijk zijn voor iets als wat ze doen verkeerd is volgens sommige morele of wettelijke normen.

In dit werk stellen we logische systemen voor om een begin te maken aan het beantwoorden van de bovenstaande vragen. Kenmerkend voor onze bijdrage is de centrale rol die de noties van agentschap en actie spelen in de formele kaders die we voorstellen. Het proefschrift is opgedeeld in twee delen. In deel I ontwikkelen we logica's om te redeneren over causale verantwoordelijkheid en om de interactie tussen agentschap en contrafactisch redeneren te analyseren. Ons uitgangspunt is één van de meest prominente logica's van agentschap in de filosofische literatuur: STIT-logica (de logica van *ervoor zorgen dat*). We beginnen, in hoofd-

stuk 3, met het verfijnen van STIT-logica om er echte causale noties in op te nemen. We formaliseren drie sleuteltesten om causale verantwoordelijkheid toe te kennen, wat aanleiding geeft tot drie overeenkomstige STIT-operators en gebruiken ze om individuele- en groepsverantwoordelijkheid te analyseren in een aantal voorbeelden. Hoofdstuk 4 breidt het raamwerk uit dat in Hoofdstuk 3 is geïntroduceerd en combineert het met een logica van contrafactische implicaties. We presenteren drie nieuwe vormen van STIT-semantiek voor contrafactische implicaties en bespreken belangrijke filosofische en logische gevolgen die hieruit voortvloeien. In Hoofdstuk 5 gebruiken we technieken uit STIT-logica, epistemische logica en onderwerp-semantiek om een model te ontwikkelen van de mentale activiteit die ten grondslag ligt aan de evaluatie van contrafactische uitspraken, namelijk verbeelding geïnterpreteerd als realiteitsgerichte geestelijke simulatie. We evalueren wat de logica van een dergelijke activiteit is, wat de vrijwillige en onvrijwillige componenten ervan zijn en, gerelateerd, hoe deze kennis genereert.

In deel II bestuderen we deontische logica om de verschillende manieren waarop het doen van iets 'fout' kan zijn te analyseren. Het belangrijkste kenmerk van de logische systemen die in dit deel worden ontwikkeld, is dat ze zijn gebaseerd op dynamische logica, d.w.z. logica's die *acties* modelleren als overgangen van een begintoestand (of model) naar een eindtoestand (of model). Hoofdstuk 6 presenteert een dynamische deontische logica die wordt gekenmerkt door de noties van idealiteit en optimaliteit. We gebruiken deze begrippen om een fijnmazige deontische classificatie te geven van toestanden, acties en opeenvolgingen van acties en om deontische operators te definiëren die zogenaamde feitelijke voorschriften uitdrukken — voorschriften die gevoelig zijn voor wat er, gegeven de omstandigheden, daadwerkelijk gedaan kan worden. Werkelijke voorschriften zijn van groot belang in situaties waarin agenten niet anders kunnen dan bepaalde normen overtreden. Hoofdstuk 7 weidt uit over een hoofdcategorie van dergelijke situaties, namelijk situaties die het gevolg zijn van een normatief conflict. Door gebruik te maken van de middelen van expliciete modale logica en dynamische epistemische logica, ontwerpen we een raamwerk om de dynamiek die aanleiding geeft tot een conflict te modelleren. We laten zien hoe het resulterende systeem kan worden gebruikt om de agenten die een conflict hebben veroorzaakt bij te houden en om onderscheidende aspecten van gevallen van gewetensbezwaren en burgerlijke ongehoorzaamheid vast te leggen.

Summary

Where Responsibility Takes You Logics of Agency, Counterfactuals and Norms

This dissertation studies logical systems merging insights from logics of agency, counterfactuals, and norms. The aim is to develop tools to address three general questions related to a formal analysis of *causal responsibily* (i.e., responsibility for what happened, regardless of one's intentions or beliefs): How can we model the agency of individuals and groups in causing certain results in complex multiagent scenarios? What are the logical properties and epistemic value of counterfactuals concerning what can be, or could have been, done in the course of time? Which rules govern normative reasoning? The first question derives from the fact that agents are only responsible for what they caused. The second arises because causal responsibility is typically determined by considering what would have happened had the relevant agents acted differently. The third stems from the fact that agents are responsible for something only if what they did was wrong according to some moral or legal norms.

In this work, we propose logical systems to begin to answer the aforementioned questions. A characterizing feature of our contribution is the central role played by the notions of agency and action in the formal frameworks we advance. The thesis is organized as follows.

In Part I, we develop logics to reason about causal responsibility and to analyze the interaction between agency and counterfactual reasoning. Our point of departure is one of the most prominent logics of agency in the philosophical literature, namely STIT logic (the logic of *seeing to it that*). We start, in Chapter 3, by refining STIT logic in order to include genuinely causal notions in it. We formalize three key tests to ascribe causal responsibility, giving rise to three corresponding STIT operators, and use them to analyze individual and group responsibility in a number of examples. Chapter 4 extends the framework introduced in Chapter 3 and combines it with a logic of counterfactuals. We present three new STIT

semantics for counterfactuals and discuss important philosophical and logical implications deriving from them. In Chapter 5, we use techniques from STIT logic, epistemic logic, and subject matter semantics to advance a model of the mental activity that underlies the evaluation of counterfactual statements, namely imagination intended as reality oriented mental simulation. We consider what the logic of such activity is, what its voluntary and involuntary components are, and, relatedly, how it generates knowledge.

In Part II, we study deontic logics to analyze the senses in which doing something can be "wrong." The hallmark of the logical systems developed in this part is that they are based on dynamic logics, i.e., logics modeling *actions* as transitions from an initial-state (or model) to an end-state (or model). Chapter 6 presents a dynamic deontic logic characterized by both a notion of ideality and a notion of optimality. We use these notions to provide a fine-grained deontic classification of states, actions, and sequences of actions and to define deontic operators expressing so-called actual prescriptions – prescriptions that are sensitive to what can actually be done, given the circumstances. Actual prescriptions are of the greatest importance in situations in which the agents cannot avoid violating some norms. Chapter 7 zooms in on a main category of such situations, namely those resulting from the presence of a normative conflict. By relying on the resources of explicit modal logic and dynamic epistemic logic, we design a framework to model the dynamics that gives rise to a conflict. We show how the resulting system can be used to keep track of the agents who generated a conflict and to capture distinctive aspects of cases of conscientious objection and civil disobedience.

Titles in the ILLC Dissertation Series:

ILLC DS-2009-01: **Jakub Szymanik**
Quantifiers in TIME and SPACE. Computational Complexity of Generalized Quantifiers in Natural Language

ILLC DS-2009-02: **Hartmut Fitz**
Neural Syntax

ILLC DS-2009-03: **Brian Thomas Semmes**
A Game for the Borel Functions

ILLC DS-2009-04: **Sara L. Uckelman**
Modalities in Medieval Logic

ILLC DS-2009-05: **Andreas Witzel**
Knowledge and Games: Theory and Implementation

ILLC DS-2009-06: **Chantal Bax**
Subjectivity after Wittgenstein. Wittgenstein's embodied and embedded subject and the debate about the death of man.

ILLC DS-2009-07: **Kata Balogh**
Theme with Variations. A Context-based Analysis of Focus

ILLC DS-2009-08: **Tomohiro Hoshi**
Epistemic Dynamics and Protocol Information

ILLC DS-2009-09: **Olivia Ladinig**
Temporal expectations and their violations

ILLC DS-2009-10: **Tikitu de Jager**
"Now that you mention it, I wonder...": Awareness, Attention, Assumption

ILLC DS-2009-11: **Michael Franke**
Signal to Act: Game Theory in Pragmatics

ILLC DS-2009-12: **Joel Uckelman**
More Than the Sum of Its Parts: Compact Preference Representation Over Combinatorial Domains

ILLC DS-2009-13: **Stefan Bold**
Cardinals as Ultrapowers. A Canonical Measure Analysis under the Axiom of Determinacy.

ILLC DS-2010-01: **Reut Tsarfaty**
Relational-Realizational Parsing

ILLC DS-2010-02: **Jonathan Zvesper**
Playing with Information

ILLC DS-2010-03: **Cédric Dégremont**
The Temporal Mind. Observations on the logic of belief change in interactive systems

ILLC DS-2010-04: **Daisuke Ikegami**
Games in Set Theory and Logic

ILLC DS-2010-05: **Jarmo Kontinen**
Coherence and Complexity in Fragments of Dependence Logic

ILLC DS-2010-06: **Yanjing Wang**
Epistemic Modelling and Protocol Dynamics

ILLC DS-2010-07: **Marc Staudacher**
Use theories of meaning between conventions and social norms

ILLC DS-2010-08: **Amélie Gheerbrant**
Fixed-Point Logics on Trees

ILLC DS-2010-09: **Gaëlle Fontaine**
Modal Fixpoint Logic: Some Model Theoretic Questions

ILLC DS-2010-10: **Jacob Vosmaer**
Logic, Algebra and Topology. Investigations into canonical extensions, duality theory and point-free topology.

ILLC DS-2010-11: **Nina Gierasimczuk**
Knowing One's Limits. Logical Analysis of Inductive Inference

ILLC DS-2010-12: **Martin Mose Bentzen**
Stit, Iit, and Deontic Logic for Action Types

ILLC DS-2011-01: **Wouter M. Koolen**
Combining Strategies Efficiently: High-Quality Decisions from Conflicting Advice

ILLC DS-2011-02: **Fernando Raymundo Velazquez-Quesada**
Small steps in dynamics of information

ILLC DS-2011-03: **Marijn Koolen**
The Meaning of Structure: the Value of Link Evidence for Information Retrieval

ILLC DS-2011-04: **Junte Zhang**
System Evaluation of Archival Description and Access

ILLC DS-2011-05: **Lauri Keskinen**
Characterizing All Models in Infinite Cardinalities

ILLC DS-2011-06: **Rianne Kaptein**
Effective Focused Retrieval by Exploiting Query Context and Document Structure

ILLC DS-2011-07: **Jop Briët**
Grothendieck Inequalities, Nonlocal Games and Optimization

ILLC DS-2011-08: **Stefan Minica**
Dynamic Logic of Questions

ILLC DS-2011-09: **Raul Andres Leal**
Modalities Through the Looking Glass: A study on coalgebraic modal logic and their applications

ILLC DS-2011-10: **Lena Kurzen**
Complexity in Interaction

ILLC DS-2011-11: **Gideon Borensztajn**
The neural basis of structure in language

ILLC DS-2012-01: **Federico Sangati**
Decomposing and Regenerating Syntactic Trees

ILLC DS-2012-02: **Markos Mylonakis**
Learning the Latent Structure of Translation

ILLC DS-2012-03: **Edgar José Andrade Lotero**
Models of Language: Towards a practice-based account of information in natural language

ILLC DS-2012-04: **Yurii Khomskii**
Regularity Properties and Definability in the Real Number Continuum: idealized forcing, polarized partitions, Hausdorff gaps and mad families in the projective hierarchy.

ILLC DS-2012-05: **David García Soriano**
Query-Efficient Computation in Property Testing and Learning Theory

ILLC DS-2012-06: **Dimitris Gakis**
Contextual Metaphilosophy - The Case of Wittgenstein

ILLC DS-2012-07: **Pietro Galliani**
The Dynamics of Imperfect Information

ILLC DS-2012-08: **Umberto Grandi**
Binary Aggregation with Integrity Constraints

ILLC DS-2012-09: **Wesley Halcrow Holliday**
Knowing What Follows: Epistemic Closure and Epistemic Logic

ILLC DS-2012-10: **Jeremy Meyers**
Locations, Bodies, and Sets: A model theoretic investigation into nominalistic mereologies

ILLC DS-2012-11: **Floor Sietsma**
Logics of Communication and Knowledge

ILLC DS-2012-12: **Joris Dormans**
Engineering emergence: applied theory for game design

ILLC DS-2013-01: **Simon Pauw**
Size Matters: Grounding Quantifiers in Spatial Perception

ILLC DS-2013-02: **Virginie Fiutek**
Playing with Knowledge and Belief

ILLC DS-2013-03: **Giannicola Scarpa**
Quantum entanglement in non-local games, graph parameters and zero-error information theory

ILLC DS-2014-01: **Machiel Keestra**
Sculpting the Space of Actions. Explaining Human Action by Integrating Intentions and Mechanisms

ILLC DS-2014-02: **Thomas Icard**
The Algorithmic Mind: A Study of Inference in Action

ILLC DS-2014-03: **Harald A. Bastiaanse**
Very, Many, Small, Penguins

ILLC DS-2014-04: **Ben Rodenhäuser**
A Matter of Trust: Dynamic Attitudes in Epistemic Logic

ILLC DS-2015-01: **María Inés Crespo**
Affecting Meaning. Subjectivity and evaluativity in gradable adjectives.

ILLC DS-2015-02: **Mathias Winther Madsen**
The Kid, the Clerk, and the Gambler - Critical Studies in Statistics and Cognitive Science

ILLC DS-2015-03: **Shengyang Zhong**
Orthogonality and Quantum Geometry: Towards a Relational Reconstruction of Quantum Theory

ILLC DS-2015-04: **Sumit Sourabh**
Correspondence and Canonicity in Non-Classical Logic

ILLC DS-2015-05: **Facundo Carreiro**
Fragments of Fixpoint Logics: Automata and Expressiveness

ILLC DS-2016-01: **Ivano A. Ciardelli**
Questions in Logic

ILLC DS-2016-02: **Zoé Christoff**
Dynamic Logics of Networks: Information Flow and the Spread of Opinion

ILLC DS-2016-03: **Fleur Leonie Bouwer**
What do we need to hear a beat? The influence of attention, musical abilities, and accents on the perception of metrical rhythm

ILLC DS-2016-04: **Johannes Marti**
Interpreting Linguistic Behavior with Possible World Models

ILLC DS-2016-05: **Phong Lê**
Learning Vector Representations for Sentences - The Recursive Deep Learning Approach

ILLC DS-2016-06: **Gideon Maillette de Buy Wenniger**
Aligning the Foundations of Hierarchical Statistical Machine Translation

ILLC DS-2016-07: **Andreas van Cranenburgh**
Rich Statistical Parsing and Literary Language

ILLC DS-2016-08: **Florian Speelman**
Position-based Quantum Cryptography and Catalytic Computation

ILLC DS-2016-09: **Teresa Piovesan**
Quantum entanglement: insights via graph parameters and conic optimization

ILLC DS-2016-10: **Paula Henk**
Nonstandard Provability for Peano Arithmetic. A Modal Perspective

ILLC DS-2017-01: **Paolo Galeazzi**
Play Without Regret

ILLC DS-2017-02: **Riccardo Pinosio**
The Logic of Kant's Temporal Continuum

ILLC DS-2017-03: **Matthijs Westera**
Exhaustivity and intonation: a unified theory

ILLC DS-2017-04: **Giovanni Cinà**
Categories for the working modal logician

ILLC DS-2017-05: **Shane Noah Steinert-Threlkeld**
Communication and Computation: New Questions About Compositionality

ILLC DS-2017-06: **Peter Hawke**
The Problem of Epistemic Relevance

ILLC DS-2017-07: **Aybüke Özgün**
Evidence in Epistemic Logic: A Topological Perspective

ILLC DS-2017-08: **Raquel Garrido Alhama**
Computational Modelling of Artificial Language Learning: Retention, Recognition & Recurrence

ILLC DS-2017-09: **Miloš Stanojević**
Permutation Forests for Modeling Word Order in Machine Translation

ILLC DS-2018-01: **Berit Janssen**
Retained or Lost in Transmission? Analyzing and Predicting Stability in Dutch Folk Songs

ILLC DS-2018-02: **Hugo Huurdeman**
Supporting the Complex Dynamics of the Information Seeking Process

ILLC DS-2018-03: **Corina Koolen**
Reading beyond the female: The relationship between perception of author gender and literary quality

ILLC DS-2018-04: **Jelle Bruineberg**
Anticipating Affordances: Intentionality in self-organizing brain-body-environment systems

ILLC DS-2018-05: **Joachim Daiber**
Typologically Robust Statistical Machine Translation: Understanding and Exploiting Differences and Similarities Between Languages in Machine Translation

ILLC DS-2018-06: **Thomas Brochhagen**
Signaling under Uncertainty

ILLC DS-2018-07: **Julian Schlöder**
Assertion and Rejection

ILLC DS-2018-08: **Srinivasan Arunachalam**
Quantum Algorithms and Learning Theory

ILLC DS-2018-09: **Hugo de Holanda Cunha Nobrega**
Games for functions: Baire classes, Weihrauch degrees, transfinite computations, and ranks

ILLC DS-2018-10: **Chenwei Shi**
Reason to Believe

ILLC DS-2018-11: **Malvin Gattinger**
New Directions in Model Checking Dynamic Epistemic Logic

ILLC DS-2018-12: **Julia Ilin**
Filtration Revisited: Lattices of Stable Non-Classical Logics

ILLC DS-2018-13: **Jeroen Zuiddam**
Algebraic complexity, asymptotic spectra and entanglement polytopes

ILLC DS-2019-01: **Carlos Vaquero**
What Makes A Performer Unique? Idiosyncrasies and commonalities in expressive music performance

ILLC DS-2019-02: **Jort Bergfeld**
Quantum logics for expressing and proving the correctness of quantum programs

ILLC DS-2019-03: **András Gilyén**
Quantum Singular Value Transformation & Its Algorithmic Applications

ILLC DS-2019-04: **Lorenzo Galeotti**
The theory of the generalised real numbers and other topics in logic

ILLC DS-2019-05: **Nadine Theiler**
Taking a unified perspective: Resolutions and highlighting in the semantics of attitudes and particles

ILLC DS-2019-06: **Peter T.S. van der Gulik**
Considerations in Evolutionary Biochemistry

ILLC DS-2019-07: **Frederik Möllerström Lauridsen**
Cuts and Completions: Algebraic aspects of structural proof theory

ILLC DS-2020-01: **Mostafa Dehghani**
Learning with Imperfect Supervision for Language Understanding

ILLC DS-2020-02: **Koen Groenland**
Quantum protocols for few-qubit devices

ILLC DS-2020-03: **Jouke Witteveen**
Parameterized Analysis of Complexity

ILLC DS-2020-04: **Joran van Apeldoorn**
A Quantum View on Convex Optimization

ILLC DS-2020-05: **Tom Bannink**
Quantum and stochastic processes

ILLC DS-2020-06: **Dieuwke Hupkes**
Hierarchy and interpretability in neural models of language processing

ILLC DS-2020-07: **Ana Lucia Vargas Sandoval**
On the Path to the Truth: Logical & Computational Aspects of Learning

ILLC DS-2020-08: **Philip Schulz**
Latent Variable Models for Machine Translation and How to Learn Them

ILLC DS-2020-09: **Jasmijn Bastings**
A Tale of Two Sequences: Interpretable and Linguistically-Informed Deep Learning for Natural Language Processing

ILLC DS-2020-10: **Arnold Kochari**
Perceiving and communicating magnitudes: Behavioral and electrophysiological studies

ILLC DS-2020-11: **Marco Del Tredici**
Linguistic Variation in Online Communities: A Computational Perspective

ILLC DS-2020-12: **Bastiaan van der Weij**
Experienced listeners: Modeling the influence of long-term musical exposure on rhythm perception

ILLC DS-2020-13: **Thom van Gessel**
Questions in Context

ILLC DS-2020-14: **Gianluca Grilletti**
Questions & Quantification: A study of first order inquisitive logic

ILLC DS-2020-15: **Tom Schoonen**
Tales of Similarity and Imagination. A modest epistemology of possibility