Logics of Responsibility

Aldo Iván Ramírez Abarca

θπ

Logics of Responsibility

Logica's van Verantwoordelijkheid

(met een samenvatting in het Nederlands)

Proefschrift

ter verkrijging van de graad van doctor aan de
Universiteit Utrecht
op gezag van de
rector magnificus, prof.dr. H.R.B.M. Kummeling,
ingevolge het besluit van het college voor promoties
in het openbaar te verdedigen op

woensdag 7 juni 2023 des ochtends te 10.15 uur

door

Aldo Iván Ramírez Abarca

geboren op 5 maart 1987
te Mexico City, México

Promotor:
Prof. dr. J. M. Broersen

Co-promotor:
Dr. H. W. A. Duijf

Beoordelingscommissie:
Dr. N. A. Alechina
Prof. dr. D. Cohnitz (voorzitter)
Dr. E. Lorini
Prof. dr. O. Roy
Prof. dr. S. J. L. Smets

Cover illustration by Moriz Oberberger.

ISBN: 978-94-93253-07-0

For Karen and Goya,
for my cicciolini,
and for my brothers.

Acknowledgments

This thesis is the result of the will and work of many people. I will not lie. It damn near did not came to be, and I wish to thank all those who were involved in making it happen.

Jan, your supervision throughout these years was all that someone can wish for, and then some more. The support was essential and immense. Both as a professional and as a person, you are a big role model for me, and I thank you for that. Hein, all your contributions and careful observations, as well as your meticulous review, turned a mess of a manuscript into something readable. I'm truly grateful for the commitment that you put into my crossing the finish line.

Karen, Beto, and Goya, you know that I know that you know that, without you, none of it could have been possible. Luis Efrén and Lorena, thank you for pushing hope when push came to shove.

Thank you, my cicciolini, for being the family that I needed. Anits, Moriz, Oli, Dimi, Xari, Despo, and Pau, these pages are as much mine as they are yours, since what you always brought to the table brought them to life. Anits, an everlasting grace and an everlasting flame for the light; Moriz, the bombs of laughter and the delight of playing the sport; Oli, the waves of bravery, the beauty, and the pull of the words; Dimi, the joy of stargazers and the strength of a sun; Xari, the elegance and the grit, the heart's music within; Despo, the zest and the color in all those stories of spirit and drive; Pau, the movement and the smiles and their "swirls all around."

Thank you, my brothers, for shouldering me to put up the fight. Adu, Kostis, and Bujar, you picked me up and put me on track. Chucky, you eased the load with the power of a perpetually raised fist. Raúl, Fern, and Aless, you always pulled me into truth and courage and the will to not back down. Nigel and Johannes, you played the parts so that the song could be sung. Salva, you showed me how to face things in questa 'hermosa y formosa' vita: guidati dalla ragione ma spinti dal cuore.

I have said that the thesis came from the will and work of many people. So thank you, Mariau, Fer Lavín, Alonso mi primo, Estebitan, Jerome, Marianits, Gauti, Jonathan, Derek, Kriss, Lemmy, Niko, Aggele, Emi, Dimitri, Lollo, Anjum, Anchi, Karo, Anna, Danilo, Aybuke, Ilaria, Marieke, Lara, Izzy, Coco, Luca, Eleni, Jelena, Ela, Saskia, Nynke, Tatiana, Baltag, Sonja, Suzie Q, Sophie, Alexandra, Maxim, Rebecca, Kees, Els, Juan Paco, Ana, Anita, Juan Champ, Melissa, Tocino, Ale Superman, Chunga, Petrone, Pablito, Pumps, Cacho, Joni, Betan, Ceci, Dersu, Gabbits, Rocha, Montse, Iames, James, Eileen, Stephen, Marianita, Mich, Marianne, Mona, Adri, Rick, Cécile, Maris, Renata de Troya, Memis, Paco, Edith, Paquito, Dani, and Sebas. I thank you all.

Special words of thanks: to Adu and Kostis for being paranymphs—and to Anits, Buki, Dimi, Despo, and Oli for helping them at the job; to Moriz for the magnificent design of the cover; to Sophie for the Samenvatting; to Suzie Q for her warmth and the constant, crucial help in all office matters; and to Sonja, Natasha, Olivier, Emiliano, and Daniel for reading and assessing the thesis.

Contents

1

Introduction

Any comprehension is only the placing of the essence of life under the laws of reason.

Leo Tolstoi, *War and Peace*

But men love abstract reasoning and neat systematization so much that they think nothing of distorting the truth, closing their eyes and ears to evidence to preserve their logical constructions.

Fyodor Dostoevsky, *Notes from the Underground*

Consider the act of parricide. Its infamy is so widespread that Fyodor Dostoevsky used the theme of killing one's own parent to write a novel that would explore the depths of human morality. Dostoevsky's *Brothers Karamazov* tells the story of Mitya Karamazov, a tempestuous man accused of having murdered his father, Fyodor Karamazov. The question of whether Mitya was in fact responsible is what drives the arches of the novel's major characters, each confronted with—and somewhat tortured by—the urge of figuring out who should be punished for the killing. For instance, Mitya's brother—Ivan—loses his mind trying to decide his own degree of responsibility. Likewise, the trial of Mitya forces the other characters—and us readers with them—to reflect on the significance of a defendant's mental states in the ascription of culpability.

Now consider the other side of the spectrum. Think of Irena Sendler, Eugene Lazowski, Nicholas Winton, or Oskar Schindler. These people devised efficient methods to save Jewish citizens from the Nazi during the Holocaust.[1] Most of us deem their deeds commendable if not heroic, and all were decorated for their humanitarianism. Although none operated individually, history has come to see them as responsible for having saved many lives in World War II.

These two examples say something important about human societies: some acts unavoidably call for the investigation of the people responsible for performing them.[2] In this thesis, I study the concept of *responsibility.* Addressing such a complex phenomenon is no easy task, and over the years a wide variety of approaches have taken up the challenge. The one adopted here is of a *formal* nature. This means that I use Mathematics and Logic to analyze the immensely broad, immensely pervasive notion of responsibility. To clarify the targets (and limits) of my work, the present introductory chapter is devoted to discussing the following points:

(a) The thesis's main topics: responsibility, agency, knowledge, belief, intentions, and obligations.

(b) The goal underlying my treatment of the main topics: creating a theory of responsibility.

(c) The tools and methodology used: logics with sound and complete proof systems.

(d) My motivation: to help in the design and verification of symbolic ethical AI.

[1] *Irena Sendler* was a Polish nurse that smuggled Jewish children out of the Warsaw Ghetto and sheltered them with local families or in orphanages. Although she was arrested and tortured by the Gestapo, she never revealed the children's location. *Eugene Lazowski* was a Polish doctor who inoculated the population of the town of Rozwadów with a strain of bacteria that made them test positive for typhus without actually having the disease. Thus, he faked an outbreak of typhus in said town and led the Germans to enforcing a quarantine of its people. This saved around 8000 persons from being sent to concentration camps. *Nicholas Winton* was a British banker that led an operation to rescue 669 children, mostly Jewish, from Czechoslovakia right before the start of World War II. He allocated the children with foster families and arranged for their safe travel to the UK. *Oskar Schindler* was a German company-owner that saved the lives of 1200 Jews in occupied Poland. These people were employees at enamelware factories that he owned, and he managed to save them by bribing SS officials in order to prevent their execution.

[2] To ascribe responsibility of an act to a person often implies one of two social intents: the intent to reprimand the person—if the act is generally condemned—or the intent to honor them—if the act is appreciated. When someone is held responsible for a deplorable act, that person is thought of as being blameworthy, and when someone is held responsible for a commendable act, they are thought of as being praiseworthy. Blameworthy people are sanctioned for whatever activity they engaged in that led to their being blamed, and praiseworthy people are decorated accordingly. This is part of a scheme, common to most human societies, where some kinds of actions are discouraged—by the sanctions given to blameworthy actors—and some kinds of actions are promoted—by the decorations awarded to praiseworthy actors.

(e) The overall position of the thesis in the academic literature.

In short, I aim to build a formal theory of responsibility. The main tool used toward this aim is Logic—and, more specifically, modal logic. The underlying motivation is to provide theoretical foundations for symbolic techniques in the development of ethical AI. Thus, this work means a contribution to the research fields known as formal philosophy and symbolic AI.

1.1 Components of Responsibility

As for point *(a)*, the main topic of this thesis is responsibility. Being an intricate concept, I opt to work with the following operational definition:

> ***Responsibility:*** *a relation between the agents and the states of affairs of an environment, such that an agent is responsible for a state of affairs iff the agent's degree of involvement in the realization of that state of affairs warrants blame or praise (in light of a given normative system).*[3]

As such, I focus on what the literature refers to as *backward-looking responsibility* (van de Poel, 2011), where an agent is considered to have produced a state of affairs that has already ensued and lies in the past. For instance, when a judge is trying a murder case and wants to find out who is responsible for doing the killing, the kind of responsibility considered is backward-looking responsibility. In contrast, *forward-looking responsibility* is responsibility that an agent has when expected to comply with the duty of bringing about a state of affairs in the future. When a student has to write an essay before its due date, this is the kind of responsibility that is being considered. My analysis of backward-looking responsibility, then, starts with a basic proposal of (i) *decomposing* its operational definition into specific components, and (ii) *classifying* different kinds of responsibility according to the decomposition.

As for the decomposition, consider the following list of components of responsibility:

- **Agents within an environment**: the so-called bearers of responsibility, the authors of actions, the actors. While agents are typically assumed to be

[3] A *normative system* is a system of unified norms that guides some activity—for instance, judicial law, that guides social life, or the rules of football. For the definition of the concept 'normative system' used in this thesis, the reader is referred to Raz (1999). For formal treatments of normative systems in the same tradition of my work, the reader is referred to Ågotnes, van der Hoek, Rodríguez-Aguilar, Sierra, and Wooldridge (2007) and Boella, van der Torre, and Verhagen (2006), for instance.

persons, computer programs, or robots, the environment is thought of as the physical and temporal stage on which agents interact—for instance, the world, the universe, or a specific domain of states. Thus, my interpretation of the term *agent* is best identified with what scholars have called *intentional systems* (van der Hoek & Wooldridge, 2008), meaning goal-directed entities whose activity (and presence) changes the environment and is accompanied by mental states.[4]

– **Actions**: the processes by which agents bring about states of affairs in the environment. For a given action, the states of affairs that it causes are known as the action's *effects*, and they typically change the environment in some way. In the philosophical literature, it is common to refer to the phenomenon by which agents choose and perform actions—with the accompanying mental states—as *agency* (see, for instance Schlosser, 2019). Thus, from here on, whenever I use the term 'agency,' I will be referring to this phenomenon.

– **Knowledge and belief**: mental states that concern the information available in the environment. These states are components of responsibility insofar as they explain agents' particular choices of action, and they provide justifications for situations in which an agent cannot comply with some rule of a normative system.

– **Intentions**: mental states that determine whether an action was done with the purpose of bringing about its effects. The idea is that, even if an action was consciously performed by an agent, it might still have been out of the agent's will to perform it, something that happens, for instance, if someone else forced the agent's hand. Intentions are components of responsibility insofar as agents can be excused for doing something when their actions were unintentional.

– **Ought-to-do's**: the actions that agents should perform, complying to the codes of a normative system. Such a normative system can be moral, judicial, or legal, for instance, and it is according to its tenets that agents can be either blamed or praised for bringing about some state of affairs. In other words, obligations or oughts-to-do's make up a context that provides a criterion for deciding whether an agent should be blamed or praised. The intuition is

[4] In the context of computer science and AI, Jennings (2000, p. 280) wrote: "an agent is an encapsulated computer system that is situated in some environment and that is capable of flexible, autonomous action in that environment in order to meet its design objectives."

that agents are praiseworthy only if they complied with a deontic obligation and blameworthy only if they failed to comply. I refer to such a context as the *deontic context* of responsibility.

These components are the sub-topics that I study in the thesis's chapters. The idea is to integrate them into a formalism that characterizes responsibility in clear terms and that offers a paradigm of systematic blame-or-praise assignment.

As for the classification of different kinds of responsibility, it follows Broersen's (2011a) proposal of *three categories of responsibility* (see also Duijf, 2018, Introduction), where the different categories directly correlate with the components of the decomposition:

1. *Causal responsibility*: an agent is causally responsible for a state of affairs iff the agent is the material author of the state of affairs, meaning that the agent has physically caused it.[5] The component of responsibility that this category involves is agency.

2. *Informational responsibility*: an agent is informationally responsible for a state of affairs iff the agent is the material author and it behaved knowingly, or consciously, while bringing about the state of affairs. The components that this category involves are agency, knowledge, and belief.

3. *Motivational responsibility*: an agent is motivationally responsible for a state of affairs iff the agent is the material author and it behaved knowingly and intentionally while bringing about the state of affairs. The components that this category involves are agency, knowledge, belief, and intentions.

1.2 Logic-Based Formalization of Responsibility

Let me proceed with this introductory chapter's points *(b)* and *(c)*, that respectively concern the thesis's main goal and the tools/methodology employed toward reaching this goal.

The Goal

Simply stated, the goal is to provide a logic-based framework to reason about responsibility. More specifically, I intend to use logic-based languages and models to characterize Broersen's three categories of responsibility as formulas evaluated

[5] This terminology follows criminal law's usual distinction between *material/immediate* and *intellectual/mediate* authors (see, for instance, van Sliedregt, 2012, Chapter 6).

on the models. The idea is to treat each component of responsibility (agency, knowledge, belief, intention, and obligation) as a *modality* of modal logic. For instance, if φ is a formula and α is an agent, then modalities of the form $[\alpha]\varphi$, $K_\alpha\varphi$, $I_\alpha\varphi$, and $\odot_\alpha\varphi$ will be used to express that α does φ, α knows φ, α intends φ, and α is obligated to do φ, respectively. Semantics for these modalities will be given on mathematical models designed to capture essential properties of the components—or at least properties that the philosophical literature has considered to be significant. Then, adopting a compositional approach, these basic modalities will be used to build complex formalizations of the categories of responsibility.

The Tool

As implied in the statement of my goal, the thesis's main tool is Logic. What do I mean when I say Logic? Well, I refer to a family of *syntax-and-semantics systems* and to the discipline of using this family to study a target phenomenon. In what follows, I use the term *Logic*—with capital 'L'—to refer to the full family of systems, and the term *logic*—with lowercase 'l'—to refer to a particular system within the family, where in this case I usually modify the term with an article (for example, '*a* logic of. . . ,' or '*the* logic of. . . '). Any such syntax-and-semantics system is defined as follows:

Definition 1.1. *A logic (or logic system) is a tuple of the form $\langle \mathcal{L}, \mathcal{C}\rangle$, where*

- *$\mathcal{L}$ is a set of linguistic expressions, known as the logic's* object language*. The elements of this set are known as* formulas*. These formulas are combinations of symbols according to a specific* grammar*, meaning a set of rules that determines which combinations of symbols are part of the language and which are not.*

- *$\mathcal{C}$ is a class of mathematical structures (such as sets, orders, trees, topological spaces, etc.) known as* models*, on which the formulas of the logic's object language are* interpreted *or* evaluated*. To clarify, for a model $\mathcal{M} \in \mathcal{C}$, an interpretation assigns to each formula in $\mathcal{L}$ a set of elements in $\mathcal{M}$, according to specific rules—known as* truth conditions*—that are designed to give meaning to the formulas.*

Narrowing down the description, I focus on *modal logic*. Modal logic is a subfamily of Logic where the languages include expressions that are built by applying operators to formulas in order to characterize diverse modalities. According to Garson (2021) "[a] modality is an expression (like 'necessarily or 'possibly') that is used to qualify the truth of a judgement." For instance, suppose that p denotes a proposition—with a particular truth value on a given model. One can apply the operator $\Box$ on p to express the alethic modality of p, so that $\Box p$ stands for 'p is necessary.' Then, one can test for $\Box p$'s truth value, on the same model, by checking whether the model meets a predetermined semantic rule that reflects the notion of necessity.[6]

Narrowing down the discussion even further, my study relies on a class of modal logics that is widely known as *stit theory*. Developed in a series of important papers during the late twentieth century (Belnap & Perloff, 1988; Belnap, Perloff, & Xu, 2001; Chellas, 1969, 1992; Perloff, 1991; von Kutschera, 1986), stit theory—where the acronym 'stit' stands for 'seeing to it that'—was introduced with the purpose of formalizing agency. Although Chapter 2 is dedicated in its entirety to the presentation of stit theory, it should be enlightening to already mention *why* I chose this class to formalize responsibility.

The question '*why this tool?*,' however, can encompass more than just stit theory. It also applies to my choice of modal logic and even of Logic itself. Giving an answer, then, will help me voice an intuition that is generally taken for granted in most works of applied logic: that the formalization of complex phenomena using logic-based languages and mathematical models is highly beneficial for the understanding of such phenomena. It is not about "preserving logical constructions" just for the sake of it (as one of this introduction's epigraphs says). It is about aiding in the comprehension of multifaceted, involute concepts. Thus, let me address the following three questions: '*why Logic?*,' '*why modal logic?*,' and '*why stit theory?*'

As for the first question, the fundamental reason for using Logic is that this discipline takes to its purest form the processes of (a) describing a phenomenon representation's and (b) making inferences based on these descriptions. As humans, we are faced with a wide variety of phenomena for whose understanding we find an equally wide variety of motives. In the task of understanding these phenomena, we typically first engage in a series of *representations* of them—for example, mental, linguistic, or mathematical. We then use linguistic constructs (for example, any human language) to *describe* these representations, combining representations in a compositional procedure for which certain conventions (for example, the rules of grammar for any language) prevail. The linguistic descriptions take *meaning* in those representations and in the entities that they make reference to. Thus, a description can be true or false, accurate or inaccurate, and likely or unlikely, in our physical world. On the basis of this scheme of repre-

[6] Other typical modalities in the literature on modal logic are epistemic modalities ('it is known by... that...'), temporal modalities ('it has always been the case that...,' 'it will be the case that...'), and deontic modalities ('it ought to be the case that...') (see, for instance Ballarin, 2017; Blackburn, De Rijke, & Venema, 2002).

sentation, description, and meaning, we build *inferences* about the phenomenon at hand, and we communicate said descriptions and inferences with the aim of furthering our understanding of it.[7]

It is my firm conviction that Logic, understood as the aforementioned family of syntax-and-semantics systems, is a powerful abstraction of these human processes. For a target phenomenon, (a) the representations are given by mathematical models, (b) the descriptions are given in terms of formulas of an object language, and (c) the inferences are built according both to standard laws of formal reasoning and to sets of axiomatic rules. I see three main advantages of this abstraction, and I cite them as reasons for my choosing Logic to study responsibility:

1. *Wide range of applicability*: Logic is an enormous family of systems, and each of these systems admits a high degree of flexibility in terms of (a) possible instantiations of formulas and (b) possible interpretations for the elements in the system's models. Therefore, Logic can be applied to reason about an enormous amount of phenomena.[8]

2. *Generalization & automation*: once that we have settled on the phenomenon to be studied using Logic, the apparatus of mathematical modelling (the semantic component) allows us to draw conclusions about any instance of the phenomenon. In other words, statements that are shown to be valid for a specified class of models will be statements that are appropriate for, pertinent to, or significant in any instance of the phenomenon that we manage to represent with one of the models of the class. Therefore, Logic is a tool of outstanding power for doing generalizations and for automating processes.

   To illustrate the benefits for generalization, consider the general theory of relativity. In essence, it says that we can *model* the universe as a fourth-dimensional pseudo-Riemannian manifold whose curvature is affected by

[7] To illustrate the processes mentioned in this paragraph, think of an apple. When we think of an apple, we inevitably resort to a mental representation of an entity that we associate with the word 'apple.' To better understand this apple, we describe its representation. Consider the phrases 'the apple is red' and 'the apple was green at the beginning of its cycle,' for instance. They are descriptions of the apple, they involve representations other than that of the apple, and they are built using the conventions of natural language. Relative to the particular apple that these phrases pretend to describe, we assign some meaning to them and regard them as either true or false. Now consider the phrases 'all apples are green at the beginning of their cycle' and 'all apples become red before falling from trees.' They are inferences made on the basis of the first descriptions, and they also take meaning in the form of truth or falsity.

[8] Some areas of applied logic are, for instance, philosophical logic (to reason about philosophical issues and doctrines), computational logic (to reason about computation), and any theory involving formal modelling—such as formal methods in computer science, logic-based decision- and game theory, epistemic game theory, etc.

matter according to Einstein's field equations. Therefore, any knowledge that we have about pseudo-Riemannian manifolds, about Lorentzian manifolds (i.e., pseudo-Riemannian manifolds whose metric tensor has a signature similar to the one of the spacetime continuum), about fourth-dimensional Lorentzian manifolds, and about systems of partial equations, to say the least, is bound to yield some knowledge about the universe—or at least about the representation of the universe given by general relativity.

To illustrate the benefits for automation, consider the *Monte Carlo Tree Search* algorithm for chess-playing computer programs. Here, the possibilities in a game of chess are modelled as a tree, where nodes represent board positions. The branches of the game-tree are valued according to whether they promote winning, and on the basis of those values a chess-playing program optimizes its performance by following certain rules. Thus, the rules underlying the program automate its behavior into that of a competent chess player.

3. *Clarity*: once that we have settled both on the phenomenon to be studied and on the logic to be used, one of the greatest benefits of explicitly spelling out the syntax and the semantics of this logic is the minimization of ambiguity. The specification of a logic implies the rigorous definition of three aspects: (1) the grammar of its language—that determines how to construct admissible linguistic expressions, (2) the models on which said linguistic expressions are interpreted, and (3) the truth conditions that make up a particular interpretation. In other words, we rule out (1) expressions that are not built according to the system's grammar, (2) representations that do not meet the conditions of the system's class of models, and (3) interpretations of the linguistic expressions that do not adhere to the truth conditions. All these exclusions amount to a razor-sharp definition of the logic, with little room for ambiguity regarding its components. And this directly translates into little room for ambiguity regarding the knowledge rendered about the phenomenon at hand.

   To illustrate Logic's benefits for clarity, consider once again the general theory of relativity. To decrease ambiguity in the expression of the physical consequences that a body with mass has in the universe, the physical magnitude known as 'mass' was identified with a component of a tensor (called the energy-momentum tensor). A tensor is an algebraic construct with inherent mathematical properties, and the formulation of the theory of gravitation that comes with general relativity is given precisely in the terminology of tensors. Thus, once the rigorous mathematical definitions for the terms 'manifold' and 'tensor' are established, there is little room for ambiguity as to what 'mass' means in said theory of gravitation.

As for the second question, the main reason for my choice of modal logic is, to put it bluntly, tradition. Many twentieth century philosophers realized that symbolic modal logic is suitable for characterizing the agentive qualities that I listed as components of responsibility (see, for instance, Fagin, Moses, Halpern, & Vardi, 1995; Hintikka, 1962; Perloff, 1991; Prior & Prior, 1955; von Wright, 1951). Much work has been done in this line of research, and I considered it an act of wisdom to 'stand on the shoulders of giants,' so to speak, and take advantage of the intuitions, methods, and results of applied modal logic.[9]

As for the third question, there are three main reasons for my choice of stit theory. The first is of a conceptual nature, the second revolves around the representational power of stit-theoretic semantics, and the last involves stit theory's connections with other fields:

1. *Agency is central to responsibility*: as implied by the decomposition on p. 3, agency and its effects lie at a basic level in the analysis of responsibility. Thus, my study demands a powerful—and expressive—logic of agency. Stit theory fits the bill.

2. *Simplicity, flexibility, and specificity of the models*: stit models are *simple* because they largely rely on standard modal logics—including temporal logics (see Chapter 2 for the detailed definition of stit models); they are *flexible* as they can be extended to include epistemic and deontic concepts; and they are *specific* as they can be used to model fine-grained details in scenarios of interdependent decision-making and/or multi-agent interaction (something that is illustrated by all the examples in this thesis).

3. *Connection with other theories*: it is well-known that stit theory bears a close connection, both conceptually and technically, with many other formalisms. This leads to a fruitful back-and-forth exchange of results, viewpoints, and applications. To clarify, over the last two decades researchers have shown that one can draw bridges between stit theory and theories such as *dynamic logic* (Canavotto, 2020; Herzig & Lorini, 2010; Horty, 2019; van Benthem & Pacuit, 2014), *coalition logic* (Broersen, Herzig, & Troquard, 2006b),

[9] A historical note seems suitable: Hintikka (1962) was the first to apply Kripke's modal logic in the formalization of *knowledge* and *belief*, thus giving birth to the class of the so-called epistemic and doxastic logics; von Wright (1951), for his part, adapted basic modal logic into a theory of oughts, that signified the birth both of deontic logic and of the family of logics of action. In the broadest sense, these are the fields that constitute the basis of this thesis.

alternating-time temporal logic (Broersen, Herzig, & Troquard, 2006a), and *decision- & epistemic game theory* (Abarca & Broersen, 2021a; Duijf, 2018; Tamminga, 2013), among others. Consequently, one can incorporate the concepts native to these fields into stit-theoretic analyses. For instance, due to the mentioned connections, stit theory has enough room to incorporate labelled actions and programs (from dynamic logic), the notion of ability of a group of agents (from coalition- and alternating-time temporal logic), the process of information disclosure in interdependent decision contexts (from epistemic game theory), the choice rules for decision makers (from decision theory), etc.[10] To be sure, all these ideas are highly relevant when it comes to formalizing responsibility.

The Methodology

How will I use stit theory to formalize responsibility, then? My methodology consists in the development of stit logics to reason about the interplay between the components of responsibility given on p. 3. Afterwards, these logics will be integrated into a framework that is rich enough to provide characterizations for Broersen's three categories of responsibility. If *responsibility* is the topic, *formalization* the goal, and *Logic* the tool, then the basis of the *methodology* lies in studying the properties of the components of responsibility by means of formulas that are valid (and some that are invalid) with respect to specific classes of stit models. A formula is valid with respect to a class of models iff it is true at every point in every model of the class, and it is invalid iff it is not valid. Thus, suppose that one manages to characterize the components of responsibility with modalities evaluated on (semantically appealing) stit models; then valid formulas are general truths that capture essential attributes of the modalities, and invalid formulas set down limits that distinguish the characterization from other possible formalizations.

One major contribution of this thesis is finding sound and complete proof systems for the aforementioned logics. The reason is that sound and complete systems enhance the analysis (and exposition) of valid/invalid formulas.[11] For a given modal logic, a proof system—also known as axiom system—is a collection of formulas of the logic's language such that (a) it includes a set of *axioms*, and such that (b) it is closed under so-called *rules of inference*. The axioms are a selection of formulas that capture basic qualities of the modalities, and the rules of inference allow us to prove new formulas, called *theorems*. A proof system is said to be

[10] I elaborate on these extensions in Chapter 2's Section 2.4.

[11] In the examination of a philosophical concept's main properties, the development of sound and complete proof systems is one of the most common practices of applied logic (Garson, 2021).

sound with respect to a class of models if every theorem is a valid formula on the class of models; a proof system is said to be complete with respect to a class of models if every valid formula on the class is a theorem of the proof system—i.e., provable from the axioms. In other words, for a class of models, a sound proof system generates only valid formulas, and a complete proof system generates all the valid formulas. For a given modal logic, a sound and complete proof system is known as an *axiomatization* of the logic, so that finding an axiomatization is otherwise known as *axiomatizing* the logic.

For most works of applied logic, one of the two following alternatives applies: either (i) a justification for why axiomatizations are appropriate is overlooked, because there exists a mostly silent and unchallenged agreement that "[d]emonstrating soundness & completeness of formal systems is a logician's central concern" (Garson, 2021); or (ii) authors shy away from axiomatization and soundness & completeness, most likely under a not unpopular view that in theoretical philosophy these practices have "become largely *'l'art pour l'art'*" (Enqvist, 2005, p. 1, emphasis in original). Therefore, in what follows I explicitly mention why axiomatization is important in *my* study of responsibility.

The significance of soundness & completeness results in the literature on Logic highlights the weight that logicians give to the relation between syntax and semantics. For a given logic, a soundness result provides a guarantee that whenever one applies a rule or uses an axiom from the sound proof system then the result will be a valid formula. In turn, a completeness result is a guarantee that all valid formulas are provable. Thus, a sound and complete proof system includes only truths and all the truths (that can be expressed with the logic's language) about the class of models at hand. In the discipline of Logic (and specially in modal logic), the usefulness of having sound and complete systems is considerable, and although some benefits are more akin than others to formalizing responsibility, the common thread between them is that soundness & completeness results *connect* different classes of models 'through' the proof systems.

A brief historical note seems suitable. According to Blackburn et al. (2002, Chapter 1), there are three phases in the development of modal logic: the syntactic era (1918-1959), the classical era (1959-1972), and the modern era (1972-present). The prevalence, fame, and alleged significance of soundness & completeness results arose with the transition between the syntactic era and the classical era. As its name implies, during the syntactic era the approach to doing modal logic—and this can be said about Logic in general—was syntactic. Researchers and philosophers would reason about systems in terms of their language. Starting from "intuitions as to what follows from what, at the level of language," they would formalize these intuitions by means of axioms and rules of inference (Enqvist, 2005, p. 13). Thus, even if axioms and rules were devised to characterize some philosophical notion or some phenomenon of nature, the task of studying and comparing proof systems was purely linguistic. When in the late 1950's Kripke (1959, 1963) and Hintikka (1962) presented relational semantics for modal logics (introducing the notion of *model*), this marked a game-changing event that gave birth to the classical era of modal logic (Blackburn et al., 2002). Modal logicians realized that one can differentiate proof systems using the classes of models with respect to which they are sound and complete, and a novel tradition of exploring proof systems through models ensued. Achieving soundness & completeness results became a trend thereafter, and the constant feedback between syntax and semantics began to be seen as one of the strongest suits of applied modal logic.

It is hard to dissociate the relevance of soundness & completeness results in applied modal logic from the historical evolution of ideas about syntax, semantics, and axiomatization. Even if during the syntactic era syntax was most important, the intuitions behind axioms and rules still depended on the intent of formalizing a target concept—for instance, necessity, knowledge, or belief.[12] With the dawn of the classical era, semantically driven insight started to grow. Since for any target concept some form of representation is unavoidable, logicians aimed to have a clear grasp on the relation between the representations (the models), on the one hand, and the formal linguistic expressions concerning the target concept, on the other. This led to the practice of rigorously defining classes of mathematical models so that the truth conditions for the evaluation of formulas would admit sound and complete axiomatizations. In particular, completeness results ensured that one could take the discussion about the target concept to the level of models, knowing that whatever was found to be valid would also be provable in the proof system.

A popular opinion in applied logic is that, when constructing a logic to formalize a target concept, people either (a) start with a proof system and then find a "matching interpretation that explains and motivates the system" or (b) start with a semantic structure and then "try to construct a language to reason about the imagined structure" (Enqvist, 2005, p. 13) (see also Blackburn et al., 2002; Hansson & Gärdenfors, 1973). In my view, however, there is an important reciprocity between linguistic and semantic intuitions. To illustrate how semantic considerations affect the development of proof systems, consider, for instance, the axiom $(p \to (q \to p))$ in Mendelson's (1964) axiomatization of propositional logic.

[12] The same can be said about propositional logic and first-order logic: they were axiomatized linguistically, but the properties of their respective languages—their grammar, their axioms, and their rules—were always meant to represent the process of having a correct argument for reaching a conclusion from specific premises.

This axiom clearly reflects both the idea of entailment and the truth condition for logical implication (given by the truth table for connective →). To illustrate the influence of syntax while devising semantic interpretations, consider Hintikka's (1962) use of possible-world structures to express agent α's knowledge as a modality of the form $K_\alpha\varphi$. Hintikka based $K_\alpha\varphi$ on a reflexive and transitive ordering on a set of possible worlds, precisely so that the principles of factivity of knowledge ($K_\alpha\varphi \to \varphi$) and positive introspection ($K_\alpha\varphi \to K_\alpha K_\alpha\varphi$) were validated. Showing that a proof system is sound and complete with respect to a class of models, then, is all about ensuring that the reciprocity between syntax and semantics is correct. In other words, soundness & completeness results ensure that one can safely move back and forth between syntactically and semantically driven arguments.

Without digressing further into the philosophy of soundness & completeness results, I list some important practical reasons for devoting so much work in this thesis to achieving sound and complete proof systems for the components of responsibility:

- *Clarity (once again)*: I reinstate that my aim is to study properties of the components of responsibility by means of formulas that are valid (and some that are invalid) with respect to a relevant class of stit models. Thus, the most immediate advantage of spelling out a (sound and complete) proof system is being transparent: anyone who carefully goes over the list of axioms and rules of inference is bound to get a better picture of what the fundamental tenets and inherent limitations of the logic are.

- *Soundness & completeness as tools for connecting different classes of models*: for a given proof system, consider the different classes of models such that the system is sound and complete with respect to each one of them. Formulas that are found to be valid on one of these classes will also be valid on the rest. This is very useful, since some of these classes might be (a) easier to work with than others, (b) smaller in cardinality than others, or (c) simpler than others. Therefore, one can always use the easier, smaller, or simpler classes—if any—to reason about (and also generate) valid formulas for the particular class that one is working with. An example of this can be found in epistemic logic. The typical proof system for knowledge (commonly known as **S4**) is sound and complete with respect to the class of reflexive and transitive Kripke structures. However, it is also sound and complete with respect to the class of topological spaces with a straightforward semantics known as the interior semantics. In particular, it is sound and complete with respect to the class that includes only the real line (van Bethem & Sarenac, 2004). Thus, to check whether a formula is valid on the class of *all* reflexive and transitive Kripke-structures, one can show that it is valid on the real line. Someone experienced in topology of the real line, then, is likely to prefer working with this single model over considering the whole class of reflexive and transitive orders.

- *Soundness & completeness as a measure of a proof system's appropriateness for studying a target concept*: when reasoning syntactically about the properties of a target concept (such as each component of responsibility), it is important to have a measure of what the set of theorems embodies. A soundness & completeness result makes it "possible to give a precise and natural meaning to claims that a proof system generated everything it ought to (for example, **S4** could now be claimed complete in a genuinely interesting sense: it generated all the formulas valid on reflexive and transitive frames)" (Blackburn et al., 2002, Chapter 1, p. 42, terminology adapted). In other words, if for a target concept a proof system is given without a class of models with respect to which the system is sound and complete, then it will be hard to grasp what the theorems are exactly about. Any such class of models is "a tool for analyzing proof systems: soundness results could distinguish proof systems, and completeness results could give them nice characterizations" (Blackburn et al., 2002, Chapter 1, p. 44, terminology adapted).

- *Completeness as explanation and as evaluation*: paraphrasing Enqvist (2005) and Hansson and Gärdenfors (1973), a completeness result plays two distinct—but correlated—roles: (a) it *explains* a logic, by giving an interpretation that clarifies the meaning of the proof system's theorems (see previous point), and (b) it *evaluates* the logic, in the sense that if a "strong semantic interpretation" (easy to work with, simple, elegant, etc.) is provided then this is typically seen as a "virtue of the logic" (Enqvist, 2005, p. 13, terminology adapted). These two roles are correlated, since a logic whose interpretations are more explanatory can be seen as better than one with obscure (or ambiguous) interpretations. To illustrate both roles, suppose that two logics formalizing responsibility are being compared. Their respective proof systems are similar, but whereas for one of these logics the class of models that makes its respective proof system complete includes only simple and straightforward models, for the other logic the class includes only overtly complicated models. Thus, the former logic can be seen as explaining responsibility in a simpler and more straightforward manner, so that it is better suited to formalizing responsibility than the latter.

Having thus exposed the reasons for my choice of tools and methodology, I sum up the main ideas of this subsection: with the goal of developing a formal theory of responsibility, in this dissertation I will provide stit logics for the components of responsibility—with sound and complete axiomatizations. I will then use these logics to characterize Broersen's three categories of responsibility, thus building a paradigm for systematic blame-or-praise assignment.

1.3 Responsible Intelligent Systems

Let me address this introductory chapter's point *(d)*, referring to the motivations that underlie my approach.

The Motivation

The reader might be wondering why I am invested in developing logic-based characterizations of responsibility. The main reason is that my work is part of a research project whose main objective was to create a framework for performing computational checks on responsibilities of artificially intelligent systems. An interdisciplinary enterprise lying in the intersection of artificial intelligence, legal theory, and philosophy, the *REINS* project (where 'REINS' stands for 'REsponsible Intelligent Systems') (Broersen, 2014b) was part of a growing venture in modern times: the automation of responsibility-checking in AI. The core idea of this particular project, then, was that an AI system can be modelled so that its specifications are fed to checking algorithms in the form of formulas expressing obligations, risks, abilities, plans, etc.

Both at the level of design and at the level of verification, developing formal characterizations of responsibility is a topic of increasing importance in AI (see, for instance, Arkoudas, Bringsjord, & Bello, 2005; Calegari, Ciatto, Denti, & Omicini, 2020; Coeckelbergh, 2020; Pereira & Saptawijaya, 2016). A representative of such trends, the *REINS* project was an ambitious attempt to create and implement logics of responsibility, tailored to the demand of harnessing automated/autonomous/intelligent decision-making in cases where decisions have moral implications.[13] The *REINS* project was divided in three sub-projects, whose respective goals were (1) creating formal theories of responsibility, (2) integrating

[13] In broad terms, implementing a logic refers to the practice of designing computer programs based on the syntax and the semantics of the logic. The goal is to tackle specific problems of said logic, such as, for instance, checking whether a formula is satisfied, resp. is valid, on a particular model.

these theories into logic-based approaches to normative systems (such as those provided by deontic logic), and (3) implementing the resulting formalisms, using model-checking techniques, to perform automated responsibility-checks.

The contents of this thesis belong to the efforts of sub-projects (1) and (2). Therefore, the underlying motivation refers to laying theoretical groundwork for the implementation of logics of responsibility in the development of ethical AI.

1.4 Interdisciplinarity

Finally, I address this introductory chapter's point *(e)*: the position of the thesis in the academic literature. Although much of what has been discussed so far should give the reader an idea about it, it is important to highlight that one of the main attributes of my research is its interdisciplinary character. In broad terms, this thesis lies in the intersection between *formal philosophy* and *symbolic AI*.

Pelletier (1977, p. 320, terminology adapted) wrote that *formal philosophy*'s "methodology can be seen as two-fold: to apply results of Logic in the solution of philosophical problems, and to extend the apparatus of Logic and Metamathematics so that it can comprehend under its purview more philosophical matters." From the field of formal philosophy, then, my work draws heavily on *action theory & logics of action* (for aiding in the conceptualization of stit-theoretic agency), *formal epistemology* (for modelling knowledge and belief) *deontic logic* (for incorporating logics of obligations and ought-to-do's into stit theory), and *decision- & epistemic game theory* (for analyzing interdependent decision contexts).[14]

Symbolic AI (see, for instance, Calegari et al., 2020; Flasiński, 2016; Haugeland, 1985; Nilsson & Nilsson, 1998) refers to techniques that use explicit models of knowledge and action in the development of AI systems. To clarify, an intelligent system is typically defined as an entity that can perceive its environment, can autonomously take actions toward the realization of some goal, and can enhance its performance by learning facts about the environment (see, for instance, Molina, 2020). Symbolic AI's intuition, then, is that one can create intelligent systems through the rule-based manipulation of symbols that encode knowledge and action. In other words, in symbolic AI the rendering of intelligence is governed by rules that are expressed using particular logics. Within this brand of

[14] For background texts on *action theory & logics of action*, the reader is referred to Segerberg (1992), Segerberg, Meyer, and Kracht (2017), Anscombe (1963), and Belnap et al. (2001). For *formal epistemology*, the reader is referred to Fagin et al. (1995), Halpern and Fagin (1989), Hintikka (1962), Stalnaker (2006), and van Ditmarsch, van der Hoek, Halpern, and Kooi (2015). For *deontic logic*, the reader is referred to von Wright (1951), Chellas (1980), and Horty (2001). For *decision- & epistemic game theory*, the reader is referred to Osborne and Rubinstein (1994), Perea (2012), Savage (1954), Luce and Raiffa (1957), Pacuit and Roy (2017), and Steele and Stefánsson (2016).

artificial intelligence, the paradigm that most influences my work is known as the *agent-oriented*—or *agent-based—approach* (see, for instance, Russell & Norvig, 1995; Shoham, 1993; Wooldridge & Jennings, 1995). Different types of agents are relevant in AI: intelligent agents (able to perceive, change, and learn about their environment with some objective), autonomous agents (able to control their behavior), and multi-agent systems or MAS (multiple intelligent agents that work together to achieve an objective). Thus, from agent-based symbolic AI my work draws heavily on the sub-fields known as *agentive knowledge representation* (for representing information that a computational agent can use to solve complex tasks), *agent-based modelling* (to simulate the behavior of interacting agents), and *logics for multi-agent systems* (to model a group of agents' concurrent/sequential interaction in the solution of concrete problems).[15]

All the mentioned sub-fields—from both formal philosophy and symbolic AI—share a central aspect of this thesis's methodology: the practice of *formal modelling*. According to Duijf (2018, Chapter 1, p. 9), this practice refers to building "mathematical models that are intended to represent relevant features and their interplay, relying on a back-and-forth interaction between such models and educated intuitions." Even if these models are idealizations of particular cases, they are immensely helpful in analyzing complex scenarios. In the present study, formal models are used to clarify philosophical ideas, to ground and illustrate conceptual theories, and to provide a tool that would help in the design and verification of AI.

1.5 Outline

The thesis is organised as follows:

Chapter 2: ***Agency***

This chapter presents and discusses at length *stit theory*, the powerful logic of action that is the bedrock of all other logics in the thesis. Thus, after a brief review of the main philosophical ideas about actions and agency, stit theory's basic syntax and semantics are thoroughly examined. On the one hand, the language of the logic includes atomic propositions and formulas induced by applying two modalities to them: $[\alpha]\varphi$ to express that α has seen to it that φ, and $\Box\varphi$ to express that φ is historically settled. On the other hand, the models are based on the theory of indeterministic *branching time* (Prior, 1967; Thomason, 1970). Branching time's indeterminism relies on the idea that at any moment there are different possible

[15] For detailed descriptions of these sub-fields of symbolic AI, the reader is referred to Calegari et al. (2020), Markman (2013), Helbing (2012), and van der Hoek and Wooldridge (2008).

paths in which the world evolves. Stit theory's semantics for action, then, assumes that an agent acts by constraining these possibilities to a definite subset. In this chapter I illustrate such a semantics with examples, using these examples to also explain traditional stit-theoretic notions such as ability, refraining, interdependent decisions in a group of agents, and deliberatively bringing a state of affairs that is not inevitable.

After exploring these notions, the chapter puts forward a review of the main logic-based properties of stit theory—in terms of valid/invalid formulas. To enhance the analysis, proof systems for both the full logic and its important restrictions are introduced. I scrutinize the main metalogic results (soundness, completeness, decidability) for said proof systems, and I elaborate on a prominent alternative semantics for stit theory: Kripke semantics.

Finally, the chapter reviews famous extensions of stit theory that are relevant in the context of modelling responsibility and that provide a background for the stit-theoretic logics that appear in the rest of the thesis. Each extension highlights a connection between stit theory and other disciplines in the literature, namely *propositional dynamic logic, coalition logic, alternating-time temporal logic, deontic logic, epistemic game theory,* and *epistemic logic*.

Chapter 3: ***Agency & Knowledge***

In recent years logicians in AI have argued that any comprehensive study of responsibility attribution should include a proper treatment of the interplay between agency and four kinds of agentive knowledge: *ex ante* knowledge, *ex interim* knowledge, *ex post* knowledge, and know-how (Broersen, 2011a; Duijf, 2018; Horty, 2019; Lorini, Longin, & Mayor, 2014). The first three kinds are standard from game-theoretical analyses on the stages of information disclosure across the decision making process (Pacuit & Roy, 2017), and the fourth has gained prominence both in logics of action and in deontic logic as a means to formalize ability (Herzig & Troquard, 2006; Horty & Pacuit, 2017).

The goal of this chapter, then, is to clarify previous stit-theoretic formalizations of the four kinds and to propose alternative interpretations that are more akin to my study of responsibility. For the latter purpose, the chapter introduces an expressive logic that extends atemporal basic stit theory with knowledge modalities ($K_\alpha\varphi$), modalities for 'next' and 'last' states ($X\varphi$ and $Y\varphi$), and the modality for agency of the grand coalition ($[Ags]\varphi$). As for the semantics, the formulas are evaluated on branching *discrete-time* structures. To illustrate the use of this logic in the formalization of the four kinds of knowledge, the chapter also includes and carefully dissects a complex example.

After the new characterizations of the four kinds of knowledge are discussed and compared to previous approaches, I turn my attention to the logic-based properties of the presented formalism. Thus, I introduce a sound and complete proof system for the logic. As it happens, the metalogic results of soundness and completeness are the main technical contributions of the chapter, where completeness involves a long, step-by-step procedure with novel steps that are reminiscent of some results by Schwarzentruber (2012).

The conclusion explores an extension of the presented logic with a modality for belief (whose semantics is probabilistic). Thus, the chapter serves as an analysis of the interplay between agency and knowledge/belief as components of responsibility.

Chapter 4, ***Agency, Knowledge, and Obligation***

Horty's (2001) seminal act-utilitarian stit theory of ought-to-do can be extended so that it deals with situations in which agents' knowledge plays a key role. Such an extension paves the way for a formalization of responsibility where agents can be excused for not complying with an obligation if they did not not know how to comply. Thus, this chapter is devoted to the introduction of an epistemic extension of act-utilitarian stit theory—on the road to building a nuanced theory of responsibility.

In particular, the epistemic extension leads to the disambiguation of two senses of ought-to-do: an objective one and a subjective one. While objective ought-to-do's only require the possibility of successfully complying with them, subjective ones require the possibility of both successfully and knowingly complying with them. Thus, the chapter examines two deontic modalities: $\odot_\alpha \varphi$ for agent α's objective ought-to-do's, and $\odot^S_\alpha \varphi$ for α's subjective ought-to-do's. The semantics for these modalities are based on measures of optimality for an agent's actions, where an agent objectively, resp. subjectively, ought to have seen to it that φ iff φ is an effect of all the optimal, resp. subjectively optimal, actions for that agent. Just as in Horty's (2001) approach, the measures of optimality are based on dominance solution concepts from game theory.

The starting point of my study lies in Horty's (2019) attempt to formalize an epistemic sense of ought. Such an attempt was inspired by three puzzles that pose a problem for merely extending act-utilitarian stit theory with epistemic modalities. The chapter carefully reviews Horty's (2019) proposal and builds my own as a reply to it. To illustrate my models, I use them to offer solutions to the problems posed by Horty's puzzles. To show the adequacy of my logic in the study of agency, knowledge, and obligation, I present a correspondence result between the models of Horty's solution and a limited sub-class of my own.

Furthermore, the chapter explores the logic-based properties of the logic of objective and subjective ought-to-do's in terms of valid/invalid formulas. I introduce a sound and complete proof system for objective ought-to-do's and a sound and complete proof system for subjective ought-to-do's. Once again, the metalogic results of soundness and completeness are the main technical—and novel—contributions.

The conclusion scopes out an extension of the full logic of objective and subjective ought-to-do's with a modality for belief (whose semantics is probabilistic) that leads to a new, doxastic sense of obligation based on expected-deontic-utility maximization. Thus, the chapter serves as an analysis of the interplay between agency, knowledge/belief, and ought-to-do as components of responsibility.

Chapter 5, ***Agency, Knowledge, and Intentionality***

Ascribing responsibility for a state of affairs to someone typically involves discussions of whether they acted intentionally and/or had an intention to bring about such a state. On the road to formalizing motivational responsibility, this chapter extends epistemic stit theory with intentions and with intentional actions. The concepts of intention and intentional action involve long-standing debates in philosophy. Therefore, the chapter opens with a review of usual interpretations of these concepts across the philosophical literature, paying special attention to previous logic-based approaches.

Afterwards, my proposal for a logic of intentionality is addressed. I extend epistemic stit theory with modality $I_\alpha\varphi$, meant to express that at a given point in time agent α had a present-directed intention toward the realization of φ. The semantics for this modality is based on special topologies, assigned to agents, that are added to stit-theoretic frames. For a given agent, the non-empty open sets of the associated topology are interpreted as present-directed intentions, so that if an open set is included in the set of indices where φ holds then this means that the agent had a present-directed intention toward the realization of φ. The choice of using topologies of intentions is rooted on the simplicity and straightforwardness of topological semantics. Importantly, such a semantics allowed me to deliver an intelligible characterization of intentional action in terms of present-directed intentions, namely with the conjunction $[\alpha]\varphi \wedge I_\alpha[\alpha]\varphi$. A characterization like this is an essential step for the formalization of motivational responsibility.

To illustrate my topological semantics, the chapter includes and discusses some examples. Afterwards, the formalism's logic-based properties, in terms of valid/invalid formulas, are addressed (pointing out where my theory stands with respect to the interplay between future-directed intentions, intentional action, and intention-with-which). To enhance the analysis, I present a sound and complete proof system for the developed logic. The metalogic results of soundness and completeness are the main technical contributions of the chapter, with a proof of completeness that implies drawing a bridge between relational and topological semantics (see, for instance Baltag, Bezhanishvili, Özgün, & Smets, 2016; Özgün, 2017).

The conclusion proposes a way of modelling future-directed intentions using temporal stit theory, and also comments on the extension of my framework with belief. Thus, the chapter serves as an analysis of the interplay between agency, knowledge/belief, and intentionality as components of responsibility.

Chapter 6, *Responsibility*

In this chapter I finally introduce a rich stit logic to analyze responsibility. My logic of responsibility is obtained by merging all the frameworks of the previous chapters, so that its language expresses agency, epistemic notions, intentionality, and different senses of obligation. The chapter characterizes the components of responsibility given on p. 3 using particular formulas of this language. Then, adopting a compositional approach—where complex modalities are built out of more basic ones—these characterizations of the components are used to formalize modes of responsibility, where by 'modes of responsibility' I mean combinations of sub-categories of Broersen's three categories (causal, informational, and motivational), cast against the background of particular deontic contexts.

On the one hand, the sub-categories correspond to the different versions of responsibility that one can consider according to the active and passive forms of the notion, where active responsibility concerns contributions and passive responsibility concerns omissions. Thus, with respect to a given state of affairs φ, the sub-categories are causal-active & causal-passive responsibility for φ, informational-active & informational-passive responsibility for φ, and motivational-active & motivational-passive responsibility for φ. On the other hand, the deontic context of a mode, given in terms of agents' ought-to-do's (with respect to φ), establishes whether and to what degree the combination of sub-categories involves either blameworthiness or praiseworthiness—under the premise that complying with an ought-to-do warrants praise and failing to comply warrants blame.

Using typical stit-like examples, the chapter discusses the resulting modes of responsibility, their formulaic characterizations, and the system of blame-or-praise assignment that they lead to. Afterwards, a proof system for the logic of responsibility (as well as for a technical extension of it) is presented, addressing the status of soundness & completeness results.

The conclusion (a) comments on an extension with belief and with doxastic obligations, (b) offers characterizations for the modes of *mens rea*, and (c) mentions possibilities for future work in the context of collective responsibility.

Conclusion

The thesis ends with an investigation of my work's position in symbolic ethical AI, stating possible paths for future work in this field.

Sources of the Original Material

Chapters 3–6 are based on articles that either have already been published or are currently in preparation. Below, I list these sources, chapter by chapter:

- Chapter 3 is based on my joint work with Jan Broersen (Abarca & Broersen, 2021b). Additionally, its conclusion includes a restriction of a framework that was introduced in another article with Jan Broersen (Abarca & Broersen, 2021a).
- Chapter 4 is based on an idea from an extended abstract (Broersen & Abarca, 2018a), that was later fully developed across two articles (Abarca & Broersen, 2019; Broersen & Abarca, 2018b). It is also highly influenced by my joint work with Hein Duijf, Jan Broersen, and Alexandra Kuncová (Duijf, Broersen, Kuncová, & Abarca, 2021). Additionally, its conclusion includes a restriction of a framework that was introduced in another article with Jan Broersen (Abarca & Broersen, 2021a).
- Chapter 5 is based on my work with Jan Broersen (Abarca & Broersen, 2023).
- Chapter 6 is an elaboration of a framework that was presented in an extended abstract developed with Jan Broersen (Abarca & Broersen, 2022).

1.6 A Reader's Guide

The only real prerequisite for reading this thesis is familiarity with modal logic. In particular, for all the logics in this work, the notions of validity, logical consequence, satisfiability, theoremhood, deducibility, and consistency are defined in the standard way. Likewise, I use the standard naming conventions for basic modal proof systems such as **K**, **KD**, **KD45**, and **S5**. For a background text on modal logic that includes all the concepts used here, the reader is referred to Blackburn et al. (2002).

Let me comment on the overall structure of the thesis, then. First, Chapter 2 introduces stit theory and thus lays out the basic groundwork for the remaining chapters. In principle, an expert reader can skip Chapter 2, but the examples are fun, and the clear statement of my interpretation of stit models will clarify my exact usage of stit semantics and terminology. In turn, Chapters 3–6 work as an original and coherent whole that is best read in that order. Although each of these chapters includes cross-references that would make it possible for them to be read independently, the reader is advised to at least go through their basic outlines in order. To be sure, Chapter 6 should not be read independently, because it amalgamates the previous chapters into a formal theory of responsibility.

Since each chapter comprises a wide variety of intuitions, interpretations, constructions, and conceptual arguments, each chapter includes an introduction and a conclusion. The introductions are meant to (a) motivate the discussions and sketch the chapter's goals and methodology, and to (b) help the reader keep track of the chapter's structure, providing an outline that summarizes the contents of its sections and subsections. The conclusions, for their part, are meant to (a) situate the chapter within the thesis's overall structure, and to (b) investigate potential paths for future work (rather than sum up the chapter's contents in what would feel like repetitive statements).

As mentioned above, the metalogic results for each logic in Chapters 3–6 are the thesis's main technical contributions. All the proofs of completeness involve long, step-by-step procedures that make use of Kripke semantics, canonical models, correspondence results, standard techniques of modal logic (for whose review the reader is once again referred to Blackburn et al. (2002)), and also novel techniques. For the sake of readability, every long proof in Chapters 3–6—including those for metalogic results—is relegated to an appendix.

2 Agency

> *'Time forks, perpetually, into countless futures. In one of them, I am your enemy.'*
>
> Jorge Luis Borges, *The Garden of Forking Paths*

> *Every man lives for himself, uses his freedom to achieve his personal goals, and feels with his whole being that right now he can or cannot do such-and-such action; but as soon as he does it, this action, committed at a certain moment in time, becomes irreversible and makes itself the property of history...*
>
> Leo Tolstoi, *War and Peace*

> *'Give me your hand. What's done cannot be undone.'*
>
> William Shakespeare, *Macbeth*

2.1 Introduction

Consider the *act* of parricide—yet again. At the beginning of Chapter 1 I mentioned that it is a terrible circumstance that drives people to want to find out who

[1] An expert reader should skip this chapter.

is responsible. If we ask ourselves why parricide is so terrible, we would probably answer in terms of an arbitrary division between right and wrong. This division would separate certain *actions* from others. So, before diving into discussions about the boundary of such a division (with all its implications for responsibility), let us talk about the domain in which the division arises: the domain of actions, acts, activity, actors, authors, active agents, etc.

In everyday life, the elements of this domain are very primal notions, and people rarely discuss their ontology in depth. However, *actions* and *agents* are the first elements in my list of components of responsibility (p. 3), so it is almost mandatory to spend some pages addressing the philosophical viewpoint, concerning them, that this work adopts. Virtually every logic in this thesis is an extension of the logic of action known as *stit theory*, where the acronym 'stit' stands for 'seeing to it that.' Thus, a thorough introduction to what one can call 'the conceptual building blocks of stit theory' is necessary. This chapter is devoted to such demands.

The concept of action is ever-present both in natural language and in philosophy. As for natural language, the category of verbs appears in all human languages, reflecting that the ways in which entities (designated by nouns) affect other entities is central to human understanding. As for philosophy, the starting point for my treatment of action is Aristotle's *Nicomachean Ethics III*. This famous treatise focuses on the processes, carried out by agents, that bring about changes in the world. Although Aristotle was thinking only of human agents when he presented his theory of action,[2] I use the term *agent* to refer to any author of actions. With a bit of circularity, then, I interpret actions as events that are caused—and performed—by agents, that bring about changes in the environment. Because of this definition, rather than 'action,' the word 'agency' fits better with the phenomenon lying at the core of my study.

The so-called *standard conception of action* (SCA), emerging from Anscombe's and Davidson's seminal works (Anscombe, 1963; Davidson, 1963), serves as the broadest philosophical background for my standpoint on agency. The SCA bases agency on a notion known as *intentional action*—referring to events that are performed under some intention or for some reason. An action, then, is taken as any event that is an intentional action under some description. The typical example behind this intuition is as follows: suppose that I wake up in the middle of the night and turn on a light, thus alerting a burglar that is robbing my house. If I do not know that there is a burglar there, then 'alerting the burglar' is definitely an unintentional action of mine, even if 'turning on the light' is clearly an intentional

[2] In *Nicomachean Ethics* (III, 1113b17–19), Aristotle wrote that "a human being is a first principle or the begetter of his actions as he is of children" (see Crisp, 2014, p. 45).

one. However, both actions refer to the same event, so that 'alerting the burglar' is indeed an intentional action under some description, namely the description 'turning on the light.' Thus, 'alerting the burglar' is also an action of mine: *I* alert the burglar.

It is evident that intentionality is central to the SCA. However, I will not discuss the ontology of intentions at this point, nor will I explicitly state my subscription to any particular philosophical account of what intentions are. Rather, I opt to take the category of intentions as accommodating a variety of agentive properties that—at least intuitively—motivate the choice and performance of actions. Reasons, beliefs, desires, goals, objectives, and plans, for instance, all distinguish intentional actions from unintentional ones. Thus, they provide a criterion for deciding whether or not an event is an intentional action under some description.[3]

Even with the definition provided by the SCA, agency is still an elusive concept. As pointed out by Austin (1961), 'doing an action' is a very abstract expression in the philosophical literature. To answer the question of how one is using the term *agency*, then, one should "reply by explaining its syntactics and demonstrating its semantics" (Austin, 1961, p. 28). In this chapter I will do precisely this, by introducing—and extensively discussing—the basic aspects of stit theory. An outline is included below.

- Section 2.2 explains the syntax and semantics of basic stit theory (*BST*), illustrating the models with simple examples. Afterwards, my interpretation of stit models is put forward, clarifying potential doubts regarding the use of stit theory throughout the thesis.

- Section 2.3 introduces the fundamental logic-based properties and metalogic results for temporal and atemporal *BST*, reviewing (a) the validity, resp. invalidity, of special formulas, (b) standard proof systems, (c) soundness & completeness results, and (d) alternative semantics for the logics.

- Section 2.4 reviews famous extensions of *BST* that are relevant in the context of modelling responsibility: extension with group agency, extension with action types, extension with utilities and obligations, and extension with epistemic notions. This section is mostly an exercise of foreshadowing, setting the tone for the discussions and examples that are widely adopted in succeeding chapters.

[3] It is important to remark that *intentions* are included as a quality separate from *agency* in the list of components of responsibility from Chapter 1 (p. 3). There, they represent the explicit and particular intentions that determine when a specific action is intentional and when it is not (see Chapter 5).

2.2 Stit Theory

To *see to it that* φ. This is stit theory's main modality. As in any intensional logic characterizing a philosophical concept, the semantics for the modality matter very much. In this case, the original—and most widespread—semantics was developed in a series of influential papers that appeared during the late twentieth century (Belnap & Perloff, 1988; Belnap et al., 2001; Chellas, 1969, 1992; Perloff, 1991; von Kutschera, 1986). Together, these papers gave rise to a full-fledged modal logic of agency, whose formulation answers Austin's aforementioned demands of explaining the syntactics and demonstrating the semantics of actions and agency.

The original semantics concerns agency in the world over an indeterministic conception of time known as *branching time*. A model proposed by Prior (1967) and polished by Thomason (1970), branching time's indeterminism is based on the idea that at any moment there are different possible paths in which the world evolves. Stit theory's semantics for action, then, assumes that an agent acts by constraining these possibilities to a definite subset. Directly quoting Belnap et al. (2001), "these ideas are in some part rooted in common sense. Without help, however, common sense cannot seem to pull them together into a coherent whole. One of our principal aims is to carry out that job by articulating them in a completely intelligible exact theory." Therefore, perhaps it is best to cut to the chase and introduce in all formality the original stit logic. I will henceforward refer to this logic as *basic stit theory* (*BST*), and I follow Horty and Belnap (1995) (see also Horty, 2001) for its formulation.

2.2.1 *BST*'s Syntax & Semantics

As for the syntactic aspect of *BST*, the language of the logic is defined as follows:

Definition 2.1 (Syntax of *BST*)**.** *Given a finite set Ags of agent names and a countable set of propositions P, the grammar for the formal language $\mathcal{L}$ is given by*

$$\varphi ::= p \mid \neg\varphi \mid \varphi \wedge \varphi \mid \Box\varphi \mid [\alpha]\varphi \mid \mathtt{G}\varphi \mid \mathtt{H}\varphi,$$

where p ranges over P and α ranges over Ags.

In this language, $\Box\varphi$ is meant to express the historical necessity of φ or, in other words, that φ is settled at a given moment; $\Diamond\varphi$ abbreviates $\neg\Box\neg\varphi$, and it encodes the historical possibility of φ; $[\alpha]\varphi$ expresses that 'agent α has seen to it that φ' (and $\langle\alpha\rangle\,\varphi$ abbreviates $\neg[\alpha]\neg\varphi$); $\mathtt{G}\varphi$ expresses 'φ is always *g*oing to be the case in the future'; and $\mathtt{H}\varphi$ expresses that 'φ *h*as always been the case in the past' (where $\mathtt{F}\varphi$ abbreviates $\neg\mathtt{G}\neg\varphi$ and $\mathtt{P}\varphi$ abbreviates $\neg\mathtt{H}\neg\varphi$).

As for the semantic aspect of the logic, the mathematical structures on which the formulas of $\mathcal{L}$ are evaluated are called *branching-time frames*:

Definition 2.2 (*Bt*-frames & models)**.** *A tuple* $\langle M, \sqsubset, Ags, \mathbf{Choice}\rangle$ *is called a* branching-time frame *(*bt*-frame for short) iff*

- *M is a non-empty set of* moments *and* $\sqsubset$ *is a strict partial ordering on M satisfying* no backward branching*: for all* $m, m', m'' \in M$ *such that* $m' \sqsubset m$ *and* $m'' \sqsubset m$*, either* $m' = m''$ *or* $m' \sqsubset m''$ *or* $m'' \sqsubset m'$*. Each maximal* $\sqsubset$*-chain is called a* history*, and the set of all histories is denoted by H. For* $m \in M$*,* $H_m := \{h \in H; m \in h\}$*. Tuples* $\langle m, h\rangle$ *such that* $m \in M$*,* $h \in H$*, and* $m \in h$*, are called* indices*, and the set of indices is denoted by* $I(M \times H)$*.*

  Ags is the finite set of agent names from Definition 2.1.

- **Choice** *is a function that maps each agent* α *and moment m to a partition* $\mathbf{Choice}^m_\alpha$ *of* H_m*, where the cells of such a partition represent* α*'s available choices of action at m. For* $m \in M$ *and* $h \in H_m$*,* $\mathbf{Choice}^m_\alpha(h)$ *denotes the cell that includes h. This cell represents the choice of action that* α *has performed at index* $\langle m, h\rangle$*, and I refer to it as* α*'s* current choice of action at $\langle m, h\rangle$*.* **Choice** *satisfies two conditions:*

  - (NC) No choice between undivided histories*: for all* $\alpha \in Ags$ *and* $h, h' \in H_m$*, if* $m' \in h \cap h'$ *for some* $m' \sqsupset m$*, then* $h \in L$ *iff* $h' \in L$ *for every* $L \in \mathbf{Choice}^m_\alpha$*.*
  - (IA) Independence of agency*: a function* $s : Ags \longrightarrow 2^{H_m}$ *is called a* selection function at *m if it assigns to each* α *a member of* $\mathbf{Choice}^m_\alpha$*. If* $\mathbf{Select}^m$ *denotes the set of all selection functions at m, then, for all* $m \in M$ *and* $s \in \mathbf{Select}^m$*,* $\bigcap_{\alpha \in Ags} s(\alpha) \neq \emptyset$*.*

    This condition establishes that concurrent actions by distinct agents must be independent: the choices of action of a given agent cannot affect the choices available to another (see Belnap et al., 2001; Horty & Belnap, 1995, for a discussion of this property).

*A bt-*model $\mathcal{M}$*, then, is a tuple that results from adding a valuation function* $\mathcal{V}$ *to a* bt*-frame, where* $\mathcal{V} : P \longrightarrow 2^{I(M \times H)}$ *assigns to each atomic proposition a set of indices.*

In short, *bt*-frames are tree-like structures. The nodes are called *moments* because they represent precisely that: points in time. Thus, these moments are ordered by a strict partial ordering that represents the before-after relation. To reflect the indeterminacy of the future, this ordering admits forward branching; to reflect the determinacy of the past, the ordering does not admit backward branching. Each maximal chain is known as a *history*, which represents a complete temporal evolution of the world.

The semantics for agency—relative to *bt*-models—is based on the idea that when an agent acts the agent is constraining the possible paths in which the world might evolve to those in which the agent has performed the action at hand. Therefore, the set of actions that are available to an agent at some moment is taken to be a partition of the histories passing through said moment, where each available action is identified with a cell of such a partition. In most formulations of *BST*, this modelling of agency involves the concept of *choice*—hence the name **Choice** for the function assigning to each agent and moment a partition of the histories passing through that moment. The intuition is that choices ground the actions that agents perform, so that Belnap et al. (2001, p. v), for instance, wrote:

> Before an event of choosing, there are multiple alternatives open to the agent. Furthermore, since the choice is real, so must be the alternatives, and each alternative must be as real as any other. All we can say before the moment of choice is that the agent will make one of the open choices, leaving behind the unchosen alternatives. After the choice, it is correct to say that they were once possible, but are no longer possible.[4]

As is customary in modal logic (see, for instance, Blackburn et al., 2002), *bt*-models allow us to evaluate the formulas of $\mathcal{L}$:

Definition 2.3 (Evaluation rules for stit modalities). *Let $\mathcal{M}$ be a* bt*-model. The semantics on $\mathcal{M}$ for the formulas of $\mathcal{L}$ are recursively defined by the following truth*

[4] As pointed out by Duijf (2018, Chapter 1), stit theorists are typically ambiguous when it comes to the concepts of *choice* and *action*. The idea is to represent the "possible constraints that an agent is able to exercise upon the course of events at a given moment," and that such constraints can colloquially be seen as the "actions or choices open to the agent at that moment" (Horty, 2001, p. 12). However, there is no ontological consensus about whether these forms of constraining the course of events are deliberate choices. In fact, Belnap et al. (2001, p. 18) wrote that "[s]tit theory has the advantage that it permits us to postpone attempting to fashion an ontological theory, while still advancing our grasp of some important features of action, obligation, and so on." In my experience, this room for different interpretations sometimes creates confusion among non-experts. Therefore, I settle *my* interpretation of the partitions given by the function **Choice**: the cells in such partitions are the available *choices of action* that an agent has at a given moment, where these choices of action are not necessarily deliberate choices of explicitly intentional actions. To illustrate this, when I raise my arm, I might not be consciously choosing to raise my arm, but it is still an action of mine, and therefore it will be taken as an available choice of action. I elaborate on this matter Subsection 2.2.2, where I illustrate the evaluation of $\mathcal{L}$'s formulas on *bt*-frames and explore my particular interpretation of the semantics for modality $[\alpha]\varphi$.

conditions, evaluated at index $\langle m, h\rangle$*:*

$$\begin{array}{lll}
\mathcal{M}, \langle m, h\rangle \models p & \textit{iff} & \langle m, h\rangle \in \mathcal{V}(p) \\
\mathcal{M}, \langle m, h\rangle \models \neg\varphi & \textit{iff} & \mathcal{M}, \langle m, h\rangle \not\models \varphi \\
\mathcal{M}, \langle m, h\rangle \models \varphi \wedge \psi & \textit{iff} & \mathcal{M}, \langle m, h\rangle \models \varphi \textit{ and } \mathcal{M}, \langle m, h\rangle \models \psi \\
\mathcal{M}, \langle m, h\rangle \models \Box\varphi & \textit{iff} & \textit{for all } h' \in H_m, \mathcal{M}, \langle m, h'\rangle \models \varphi \\
\mathcal{M}, \langle m, h\rangle \models [\alpha]\varphi & \textit{iff} & \textit{for all } h' \in \mathbf{Choice}^m_\alpha(h), \mathcal{M}, \langle m, h'\rangle \models \varphi \\
\mathcal{M}, \langle m, h\rangle \models \mathtt{G}\varphi & \textit{iff} & \textit{for all } m' \in h \textit{ such that } m \sqsubset m', \mathcal{M}, \langle m', h\rangle \models \varphi \\
\mathcal{M}, \langle m, h\rangle \models \mathtt{H}\varphi & \textit{iff} & \textit{for all } m' \in h \textit{ such that } m' \sqsubset m, \mathcal{M}, \langle m', h\rangle \models \varphi.
\end{array}$$

Satisfiability, validity, and general validity are defined as usual: let $\mathcal{M}$ *be a* bt*-model; then a formula* φ *of* $\mathcal{L}$ *is* satisfiable *on* $\mathcal{M}$ *iff there exists an index* $\langle m, h\rangle$ *in* $\mathcal{M}$ *such that* $\mathcal{M}, \langle m, h\rangle \models \varphi$*;* φ *is* falsifiable *or refutable on* $\mathcal{M}$ *if its negation is satisfiable;* φ *is* valid *on* $\mathcal{M}$ *iff* $\mathcal{M}, \langle m, h\rangle \models \varphi$ *for every* $\langle m, h\rangle$ *in* $\mathcal{M}$*; and* φ *is* invalid *on* $\mathcal{M}$ *iff it is not valid. For a class* C *of* bt*-models,* φ *is satisfiable—resp. invalid—with respect to* C *iff there exists a model* $\mathcal{M}$ *in* C *such that* φ *is satisfiable—resp. invalid—on* $\mathcal{M}$*;* φ *is valid with respect to* C *iff it is valid on every model of* C*. At the level of frames, a formula* φ *of* $\mathcal{L}$ *is valid on a* bt*-frame* $\mathcal{F}$ *iff it is valid on every* bt*-model based on* $\mathcal{F}$*; and* φ *is invalid on* $\mathcal{F}$ *iff it is not valid.*

2.2.2 Some Examples

It is important to comment on the implications that Definition 2.3 has for my treatment of agency. Simple examples will help me in doing so.

First, let me address the topic of the indices of evaluation, meaning those moment-history pairs of the form $\langle m, h\rangle$ such that $m \in M$ and $h \in H_m$. A history h represents a full temporal evolution of the world, so that the assumption that $m \in h$ implies that m is a moment that occurs in history h—or that h passes through m. Throughout the rest of the thesis, for a given index $\langle m, h\rangle$ in any stit model, I will say that index $\langle m, h\rangle$ is *based* on moment m and *anchored* by history h. Whenever I refer to the satisfaction of a formula at $\langle m, h\rangle$, I will use the expressions 'at index $\langle m, h\rangle$...' and 'at moment m and along history h...' interchangeably.

The fact that multiple histories can pass through a single moment is the essence of branching time's indeterminism. The use of indices as points of evaluation (for the formulas of $\mathcal{L}$), then, brings to the fore that in *bt*-models time has two fundamental dimensions:

1. The chronological dimension: for $m, m' \in M$, m' lies in the future of m iff $m \sqsubset m'$, and m' lies in the past of m iff $m' \sqsubset m$. Thus, one needs to focus on this dimension to decide whether $\mathtt{G}\varphi$ or $\mathtt{H}\varphi$ holds at some index.

2. The possible-futures dimension: for $m \in M$, this dimension is given by the set H_m—the set of all the different histories that pass through m. Once again, the histories in H_m represent different temporal evolutions of the world at m, and one has to refer to this dimension to check whether $\Box\varphi$ holds at some index.

Having introduced these two dimensions, let me illustrate instances of the evaluation of formulas involving operators $\Box$, $\mathtt{G}$, and $\mathtt{H}$. Consider the *bt*-model $\mathcal{M}$ depicted in Figure 2.1.

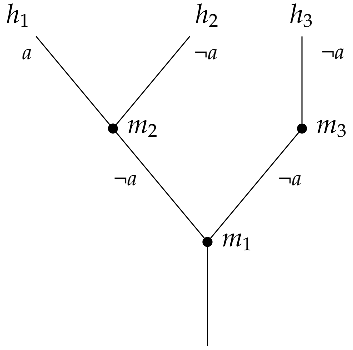


Figure 2.1: *A simple* bt-*model* $\mathcal{M}$.

Here, m_1, m_2, and m_3 are moments, and $H = \{h_1, h_2, h_3\}$. As implied by the diagram, $m_1 \sqsubset m_2$ and $m_1 \sqsubset m_3$. In all the diagrams for stit models in this thesis, the chronological dimension is reflected by the vertical axis, so that a moment that appears below another in the same history lies in the past of the first. Thus, in Figure 2.1 m_2 and m_3 lie in the future of m_1. As for the possible-futures dimension, observe that $H_{m_1} = \{h_1, h_2, h_3\}$, that $H_{m_2} = \{h_1, h_2\}$, and that $H_{m_3} = \{h_3\}$.

Let a denote the atomic proposition 'my arm is raised' in Figure 2.1. Then the diagram shows that the valuation $\mathcal{V}$ of $\mathcal{M}$ is such that $\mathcal{V}(a) = \{\langle m_2, h_1\rangle\}$. This implies that $\mathcal{M}, \langle m_1, h_1\rangle \models \Box\neg a \wedge \mathtt{F}a$: *at moment m_1 and along history h_1 my arm is not raised (and necessarily so), but at a future moment it will be the case that my arm is raised*. Now, even if at $\langle m_2, h_1\rangle$ my arm is raised, the fact that $\mathcal{M}, \langle m_2, h_2\rangle \models \neg a$ implies that it could have been otherwise. Thus, $\mathcal{M}, \langle m_2, h_1\rangle \models a \wedge \Diamond\neg a$: *at moment m_2 and along history h_1 my arm is raised, but it could have been the case that my arm were not raised*. As for other examples, consider the following: $\mathcal{M}, \langle m_3, h_3\rangle \models \Box\neg a$: *at $\langle m_3, h_3\rangle$ it was settled that my arm is not raised*; and $\mathcal{M}, \langle m_2, h_1\rangle \models a \wedge \Box\mathtt{P}\neg a$: *at $\langle m_2, h_1\rangle$ both my arm is raised and, for all histories passing through m_2, there was a moment in the past of m_2 for which my arm was not raised at the corresponding past indices.*

Observe, then, that one can use formulas to encode both the determinacy of the past and the indeterminism of the future. The determinacy of the past is reflected by the validity—with respect to the class of all *bt*-models—of formula $\mathtt{P}\varphi \to \Box\mathtt{P}\varphi$. The indeterminism of the future is reflected by the fact that $\mathtt{F}\varphi \to \Box\mathtt{F}\varphi$ is invalid (as shown by the model in Figure 2.1, where $\mathcal{M}, \langle m_1, h_1 \rangle \models \mathtt{F}a \wedge \Diamond\neg\mathtt{F}a$).

In my interpretation of stit models, each index is a 'way that things could have been.' More precisely, an index is identified with the sum of all the states of affairs—or formulas, if you will—that hold at it.[5] Both dimensions of branching time affect any index, so that if one wants to find out which formulas hold at $\langle m, h\rangle$ then one has to consider all those indices based on moments m' such that either $m' = m$ or there exists a common predecessor between m and m'. For instance, let m', m'' be moments such that $m'' \sqsubset m$ and $m'' \sqsubset m'$. If $\mathcal{M}, \langle m', h'\rangle \models \varphi$, then this has a repercussion on $\langle m, h\rangle$, since it implies that $\mathcal{M}, \langle m, h\rangle \models \mathtt{P}\Diamond\mathtt{F}\varphi$.[6]

If the chronological dimension of branching time is most natural, the possible-futures dimension is the one on which branching depends. Modality $\Diamond\varphi$ encodes historical possibility, so that $\mathcal{M}, \langle m, h\rangle \models \Diamond\varphi$ iff there exists an alternate temporal evolution of the world, given by a history h', such that φ holds at $\langle m, h'\rangle$. Now, suppose that at index $\langle m, h\rangle$ both φ holds and it is historically possible that $\neg\varphi$. In my interpretation of stit theory, this does not mean that at $\langle m, h\rangle$ someone can do something to realize $\neg\varphi$. State of affairs φ is the case, and the historical possibility of $\neg\varphi$ should be seen only as counterfactual.[7] Thus, when $\Diamond\varphi$ holds at $\langle m, h\rangle$, I will use the expression 'φ *was* historically possible at $\langle m, h\rangle$.' To clarify, at a given index either φ or $\neg\varphi$ has already been obtained, according to the valuation $\mathcal{V}$. If φ holds, $\neg\varphi$ cannot be the case anymore, unless it is thought of as a counterfactual possibility. Hence the use of the past tense, where this should not be confused with any allusion to the past operator P (see Remark 2.4).

As for the dual modality, a consistent usage of tense implies that I will say that $\mathcal{M}, \langle m, h\rangle \models \Box\varphi$ iff at $\langle m, h\rangle$ φ was settled (or historically necessary)—that is, iff φ holds at all indices based on moment m. Since φ's being settled at an index implies that φ was settled at all the other indices based on the same moment, I will henceforward use the expressions 'φ was settled at index. . . ' and 'φ was settled at moment. . . ' interchangeably.

[5] Much like in the *abstractionist* view on possible worlds (see Menzel, 2017), an index can be thought of as a *total* and *possible* state of affairs.

[6] Indeed, it can be shown that, for a given *bt*-model $\mathcal{M}$ and an index $\langle m, h\rangle$, the generated sub-model $\mathcal{M}_{\langle m,h\rangle}$ has the set $\{\langle m', h'\rangle\, ; m' = m \text{ or there exists } m'' \text{ s. t. } m'' \sqsubset m \,\&\, m'' \sqsubset m'\}$ as its domain (see Blackburn et al., 2002, Chapter 2, for the precise definitions of generated sub-models).

[7] 'Counterfactual' is a loaded term in the philosophical literature. Here, I use it in its intuitive interpretation, by which one can describe a circumstance that is not so—in 'reality,' 'actuality,' 'at the present moment,' or any other expression that indicates the actual situation—but such that it *could have been so*.

To avoid possible confusions down the line, I want to be explicit about the consequences that my interpretation—both of indices and of the evaluation of formulas on them—has on the notion of agency provided by modality $[\alpha]\varphi$. For $\alpha \in Ags$ and $m \in M$, the elements of partition $\mathbf{Choice}^m_\alpha$ represent choices of action that are available to agent α at moment m. I write 'at moment. . . ' and not 'at index. . . ' because, when a particular history $h \in H_m$ is specified, I regard that the choices that are different from $\mathbf{Choice}^m_\alpha(h)$ are not available anymore—or at least not available in any fashion other than counterfactually. At $\langle m, h\rangle$ α has both chosen and performed the choice of action represented by $\mathbf{Choice}^m_\alpha(h)$, and α cannot change its choice to any other.[8] Thus, choices of action are available at the level of moments, and only counterfactually available at the level of indices. To clarify, consider the following disambiguation:

- *At the level of moments*, an agent first arrives to a moment in branching time. There, the agent is faced with options for constraining the possible futures, given by its available choices at the moment. After possibly deliberating about these options, the agent 'decides' upon a course of action—where the quotation marks are used to emphasize that there is no explicit identification of such a decision with a conscious or intentional decision. By performing this choice at that moment, the agent indeed constrains the possible futures and brings about effects in the world.[9]

- *At the level of indices*, when a full index is specified, an agent has already deliberated, chosen, and performed one of the actions that were available at the moment on which the index is based. Furthermore, the effects of such an action have already been brought about in the world.[10]

What does it mean that α has seen to it that φ at an index, then, in my interpretation? Well, I take it that, after the performance of a choice of action

[8] It is not so much that the choice and the performance are taken to be simultaneous; rather, α has performed the action that had already been chosen, and it is by such a performance that the choice becomes evident.

[9] Each moment can be thought of as a choice scenario, with stages of choice and performance of actions. To clarify, van Benthem and Pacuit (2014) stated that there are four main stages in a choice scenario: *deliberation*, *decision*, *action*, and *observation*. "In a first deliberation stage, we analyze our options, and find optimal choices. Next, at the decision stage, we make up our mind and choose an action of our own. Then at the action stage, everyone acts publicly, and this gets observed, something that we can also model as a separate observation stage, though things happen simultaneously"(van Benthem & Pacuit, 2014, p. 309). These ideas are very important for the conceptual discussions in all the thesis's chapters.

[10] In my experience, it is tricky to pronounce oneself regarding the verb tense of the modality 'to see to it that' in stit-theoretic models. However, this is precisely what I am trying to do with this disambiguation. After presenting some examples, I discuss the verb tense that I prefer to use, in Remark 2.4.

at $\langle m, h\rangle$, consequences of such a choice—meaning effects that the action brings about in the world—are reflected by formulas φ such that $\mathcal{M}, \langle m, h\rangle \models [\alpha]\varphi$. These formulas hold at all the possible indices anchored by histories h' lying within $\mathbf{Choice}^m_\alpha(h)$. Observe that "such an action still leaves room for a good deal of variation in the course of events, and so cannot determine a unique history" (Horty, 2001, p. 12). Like Horty and Belnap (1995) (see also Horty, 2001), I interpret the histories in $\mathbf{Choice}^m_\alpha(h)$ as the possible outcomes that result from the performance of $\mathbf{Choice}^m_\alpha(h)$.

The idea that φ such that $\mathcal{M}, \langle m, h\rangle \models [\alpha]\varphi$ can be seen as an effect of $\mathbf{Choice}^m_\alpha(h)$ comes from the following reasoning. Suppose that $\mathcal{M}, \langle m, h\rangle \models \varphi$, and that there exists $h' \in \mathbf{Choice}^m_\alpha(h)$ such that $\mathcal{M}, \langle m, h'\rangle \models \neg\varphi$. By Definition 2.3, this implies that $\mathcal{M}, \langle m, h\rangle \models \neg[\alpha]\varphi$. Therefore, although φ is the case at $\langle m, h\rangle$, there is a possibility—given by $\langle m, h'\rangle$—such that α's performance of the same choice of action does not lead to φ. It is in this sense that φ should not be taken to be a material consequence, or an effect in the world, of $\mathbf{Choice}^m_\alpha(h)$.

To illustrate these claims, consider the simple *bt*-model $\mathcal{M}$ depicted in Figure 2.2.

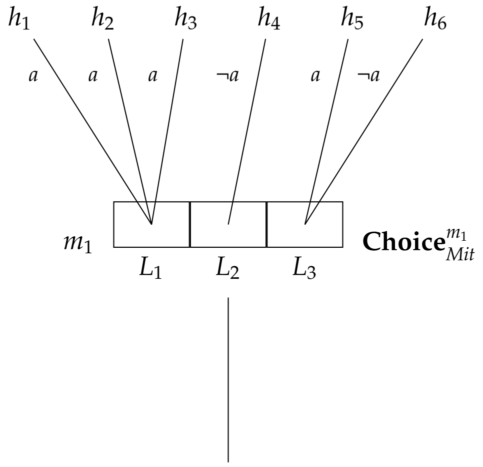


Figure 2.2: *An example of individual agency.*

Here, *Mitya* is an agent, m_1 is a moment, and $\{h_i; 1 \leq i \leq 6\}$ is the set of histories passing through m_1 (denoted by H_{m_1}). Partition $\mathbf{Choice}^m_{Mit}$ is the set $\{L_1, L_2, L_3\}$, whose elements represent choices of action that are available to *Mitya* at m_1. The diagram shows that $L_1 = \{h_1, h_2, h_3\}$, that $L_2 = \{h_4\}$, and that $L_3 = \{h_5, h_6\}$.

Let a denote the atomic proposition 'the arm of Mitya is raised' in Figure 2.2. Then the diagram shows that $\mathcal{V}(a) = \{\langle m_1, h_1\rangle, \langle m_1, h_2\rangle, \langle m_1, h_3\rangle, \langle m_1, h_5\rangle\}$.

At m_1 the available action L_1 constrains the possible futures to histories that anchor indices at which a holds. Therefore, $\mathcal{M}, \langle m_1, h_i\rangle \models [Mitya]a$ for all $1 \leq i \leq 3$: *at indices* $\langle m_1, h_1\rangle$*,* $\langle m_1, h_2\rangle$*, and* $\langle m_1, h_2\rangle$*,* Mitya *has seen to it that his arm is raised.* For this reason, a can be seen as an effect of L_1. Choice L_2, in turn, constrains the possible futures to histories anchoring indices where a does not hold, namely $\langle m_1, h_4\rangle$. Therefore, $\mathcal{M}, \langle m_1, h_4\rangle \models [Mitya]\neg a$: *at* $\langle m_1, h_4\rangle$ Mitya *has seen to it that his arm is not raised.* Finally, L_3 constrains the possible futures to histories anchoring indices such that a does not hold at all of them. Therefore, $\mathcal{M}, \langle m_1, h_i\rangle \models \neg[Mitya]a$ for all $5 \leq i \leq 6$: *at these indices* Mitya *has not seen to it that his arm is raised.* In other words, there is a temporal evolution of the world where *Mitya* has performed L_3 and his arm was not raised. Thus, a should not be seen as an effect of L_3.

Remark 2.4. In the stit-theoretic literature, people use different verb tenses for the natural-language usage of the modality *to see to it that*. While the typical presentation of the theory (Belnap & Perloff, 1988) used the present-progressive tense, with expressions of the form 'at $\langle m, h\rangle$ α is seeing to it that φ,' I follow the view of Singh (1999) and claim that $\mathcal{M}, \langle m, h\rangle \models [\alpha]\varphi$ iff at $\langle m, h\rangle$ α has just seen to it that φ (with the present-perfect tense of the expression 'to see to it'). Under this view, at $\langle m, h\rangle$ φ was achieved through α's performing of $\mathbf{Choice}^m_\alpha(h)$.[11] Three important points must be made:

1. Within a logic that admits temporal operators such as $\mathtt{G}$ and $\mathtt{H}$, using the present-perfect tense for the expression 'to see to it' might lead to controversy. If I write 'at $\langle m, h\rangle$ α has seen to it that φ,' someone might think that formula $\mathtt{H}[\alpha]\varphi$ holds at $\langle m, h\rangle$, according to which at all indices based on moments in the past of m α has seen to it that φ. This is not what I mean. My usage of the present-perfect tense—as well as the frequent use of the past tense for associated modalities—is based on the idea that the states of affairs that hold in the world at an index are definitive: they cannot be changed or be stopped from happening anymore. If $[\alpha]\varphi$ holds at $\langle m, h\rangle$, I take it that the obtaining of φ is an effect of α's performing a choice, and that φ is definitive in the world at $\langle m, h\rangle$. It is not that α is bringing about φ, or—using the still more elusive present tense—that α brings about φ. Agent

[11] Singh (1999, p. 18, emphasis in original) wrote: "Intuitively, STIT is about the actions that have just been performed. In fact, we find the progressive misleading, and believe a better gloss for STIT would be *has just seen to it that*. This gives its formal logic some characteristics different from the logics of ability or opportunity. Indeed, the concept is better understood as a form of high-level action."

α has already brought about φ, and this happened precisely at the index of evaluation. The use of the present-perfect and past tenses, then, does not involve temporal operators.

2. If φ such that $\mathcal{M}, \langle m, h \rangle \models [\alpha]\varphi$ represents an effect of the choice that α has made at $\langle m, h \rangle$, then it is evident that, in *BST*, effects are instantaneous: they ensue at the same moment at which the choice is exerted, and not in further ones. Because of this, it is not uncommon to refer to *BST* as *instantaneous stit logic* (see, for instance, Broersen, 2008a, 2011a; Payette, 2014).

3. *BST* does not include terms for actions in the logic's language. To be sure, modality $[\alpha]\varphi$ does not syntactically refer to any action that α carried out to bring about φ. The conceptualization of actions, then, occurs only at a semantic level—in terms of the choice-partitions. It is because of this that *BST* is typically seen as a "logic of actions without actions" (Horty & Pacuit, 2017; Lindström & Segerberg, 2007).

2.2.3 Ability & Refraining

With the semantics for agency and historical possibility, one can account for two important concepts that are closely related to agency: *ability* and *refraining*. These two are prominent both in the literature on stit theory and in the rest of the thesis, so let me address their characterizations in *BST*:

- *Ability:* at this point, without any explicit formalization of intentions, consciousness, or beliefs, I refer to the—very general—ability that an agent can have to cause a certain effect in the world. Therefore, it is best identified with the term *causal ability*.[12] Semantically, causal ability relies on the availability of choices of action: I am able to see to it that my hand is raised iff at this moment there exists a choice of action, available to myself, such that if I choose and perform such an action then one of its effects is that my arm is raised.

  In my interpretation of *BST*, the choices of action are available at the level of moments—and only counterfactually available at the level of indices. Thus, when a history h is specified in an index $\langle m, h \rangle$, I will say that at $\langle m, h \rangle$ an agent *was* causally able to bring about the effects of any of its available

[12] In Chapter 3 I address the notion of *ability in the epistemic sense* and its relation to know-how. In Chapter 4 the differences between causal ability and ability in the epistemic sense are brought to the fore in an investigation of deontic notions. In Chapter 6 I explore the relations between causal ability and causal responsibility, on the one hand, and ability in the epistemic sense and informational responsibility, on the other.

choices, to reinforce the idea that the agent cannot change the current choice ($\mathbf{Choice}^m_\alpha(h)$). For instance, suppose that at a given index I have seen to it that King Duncan is dead. At such an index I was also causally able to not see to it that King Duncan is dead. I was causally able in theory, but not in practice, since "what's done cannot be undone." In any case, I will use the expressions 'at index… α was able to… ' and 'at moment… α was able to… ' interchangeably. Once again, the use of the past tense does not refer to past moments; it is meant to evoke that at the level of indices states of affairs are definitive.

Formally, at $\langle m,h\rangle$ α was able to see to it that φ iff $\mathcal{M},\langle m,h\rangle \models \Diamond[\alpha]\varphi$—that is, iff it was historically possible for α to see to it that φ. Therefore, *BST*'s individual causal ability is given by a composition of the modality for historical possibility and the modality for agency. Such a composition successfully reflects the semantic idea of availability of choices at a moment: if $\Diamond[\alpha]\varphi$ holds at $\langle m,h\rangle$, then at m there is a choice of action, available to α, that brings about φ.[13]

To illustrate causal ability, consider Figure 2.2. Observe that $\mathcal{M},\langle m_1,h_1\rangle \models \Diamond[\mathit{Mitya}]\neg a$: *at $\langle m_1,h_1\rangle$ Mitya was causally able to see to it that his arm is not raised*. However, at $\langle m_1,h_1\rangle$ *Mitya* has seen to it that his arm is raised ($\mathcal{M},\langle m_1,h_1\rangle \models [\mathit{Mitya}]a$). Therefore, the most precise phrase describing this situation is the following: at $\langle m_1,h_1\rangle$ *Mitya* has seen to it that his arm is raised, and he was causally able to see to it that his arm was not raised right before making his choice of action.

- *Refraining:* broadly speaking, to refrain from doing an action can be identified with not doing the action. However, as pointed out by von Wright (1963) and by Horty (2001), to refrain intuitively implies something more than merely not doing. An example of this argument is that if I do not fly to the moon right this second it is not because I refrain from doing so; it is because it is virtually impossible for me to fly to the moon right this second.

  One can engage in a philosophical enquiry as to what refraining exactly means, and most likely the concepts of conscious choice, intentions, and in-

[13] The $\exists\forall$ pattern in the clause that defines causal ability is very reminiscent of Brown's (1988) formalization of the "can of ability" as a modality. Just as in Brown's account, the combination of modalities presented here escapes Kenny's (1976a; 1976b) well-known objections to identifying ability with a kind of possibility (see Horty & Belnap, 1995, for a reply). Indeed, the idea of using compositions of modalities to formalize attitudes closely related to agency carries over nicely to many discussions in succeeding chapters. In particular, the formalization of different modes of responsibility that I present in Chapter 6 is based on conjunctions of compositions of modalities.

tentional action will make an appearance in such discussions. Even without intentions or conscious choices in the picture, *BST* allows us to formalize two forms of refraining:

- *Von Wright's kind of refraining* (Horty, 2001, Chapter 2): at $\langle m, h\rangle$ α has von-Wright-refrained from seeing to it that φ iff $\mathcal{M}, \langle m, h\rangle \models \neg[\alpha]\varphi \wedge \Diamond[\alpha]\varphi$—that is, iff at $\langle m, h\rangle$ α has not seen to it that φ and it was possible for α to see to it that φ (where, according to our discussion on ability, the last clause means that α was causally able to see to it that φ). As will be discussed in the next subsection, I also refer to this kind of refraining as *deliberative refraining* (see Footnote 22).

- *Not doing*: at $\langle m, h\rangle$ α has refrained from seeing to it that φ iff $\mathcal{M}, \langle m, h\rangle \models \neg[\alpha]\varphi$—that is, iff at $\langle m, h\rangle$ α has not seen to it that φ.[14]

Observe that both kinds of refraining are illustrated by Figure 2.2. At $\langle m_1, h_5\rangle$ *Mitya* has both von-Wright-refrained and refrained from seeing to it that his arm is raised, even though at such an index his arm has indeed been raised (perhaps by somebody else, for instance): $\mathcal{M}, \langle m_1, h_5\rangle \models \neg[Mitya]a \wedge \Diamond[Mitya]a \wedge a$.

An alternative terminology for refraining—that I will use in my formalization of responsibility—involves the expression *omission*. Therefore, if at a given index agent α has refrained from seeing to it that φ, I take it that at such an index α has omitted seeing to it that φ, or that φ was an omission of α.[15] A related concept is that of *preventing*. In the present framework, I will say that at an index agent α has prevented φ iff $[\alpha]\neg\varphi$ holds at the index. Similarly, I will say that at an index agent α has refrained from preventing φ iff $\neg[\alpha]\neg\varphi$ holds.

[14] It is important to clarify that in some of the examples in this thesis (see Figure 6.2 and the surrounding discussion, for instance) the concept of refraining appears as a label given to a choice of action in the choice-partition $\mathbf{Choice}^m_\alpha$ (for some agent α and some moment m). This label is nothing more than a name, so that to refrain from doing any of the other available actions is explicitly interpreted as one such choice. In this sense, when at an index an agent has chosen the choice labelled as 'refrain,' the agent will have refrained from seeing to it that φ for any φ that would have been an effect of the other available choices.

[15] As will be discussed in Chapter 6, I refer to this particular type of omission, where there is no reference to the other components of responsibility (such as knowledge, beliefs, or intentions), as *causal omission*. In this context, if at an index an agent has seen to it that φ, I take it that φ was a *causal contribution* of that agent.

2.2.4 Interdependent Decision Contexts

An important feature of *BST* is that it allows us to model multi-agent situations. Up to now, I have addressed this logic focusing only on single-agent cases, but *Ags* can include various members. In the multi-agent case, it is adequate to interpret the moments of *bt*-models as *interdependent decision contexts*. Much like normal-form games, these decision contexts underlie a group of agents' interaction at some point in time. Indeed, the scheme of choice-partitions and frame condition *independence of agency* ensure that each moment of a *bt*-model looks very similar to a normal-form game that models concurrent choices of action. Since one can use *bt*-models to also represent sequential choices, then stit models can—at least in principle—also generalize aspects of the theory of extensive-form games.[16]

Let me present a simple example to illustrate *BST*'s account of concurrent and sequential multi-agent choices. Carrying on with the Dostoevsky-inspired examples, a scene from *Brothers Karamazov* might help.

Example 2.5 (*Brothers Karamazov*). *On the night of Fyrodor Karamazov's murder, Mitya Karamazov—Fyodor's son—went to Fyodor's residence in search for his lover, Grushenka, and with a burning desire to fight his father. The reason was that Grushenka was having an affair with Mitya and also with Fyodor, and the two had fallen in love with her. Mitya thought that he would find Grushenka at his father's house. Fyodor, for his part, had lately become terribly afraid of the possibility that Mitya would act on his jealousy and hurt him. Therefore, each night, including this fatal one, he would lock himself up tightly in his bedroom.*

Smerdyakov, Fyodor's bastard and lackey, was also sick of Fyodor and wanted him dead. However, Fyodor trusted him more than anyone and thus had delegated to Smerdyakov the nightly task of alerting him about whether Grushenka or Mitya had come to the household. For this, they had agreed on a 'secret code' of door knocks; if Grushenka showed up, Smerdyakov would knock on Fyodor's bedroom door one way, and if it was Mitya, he would do it some other way, letting Fyodor know who it was.

On the particular night of the murder, Smerdyakov already knew that Mitya would show up, and he had planned for Mitya to go into Fyodor's bedroom and kill the old man. Smerdyakov intended to switch the door knocks so that Fyodor would think that Grushenka had arrived and thus unlock his bedroom door. Then Mitya could go in easily and fight Fyodor. Without knowing that this was Smerdyakov's plan, since they had not agreed on anything and Mitya did not know about Smerdyakov's intentions, Mitya weighed his options when arriving to the household. He could either break into the bedroom through the window or leave. In the end, the tormented Mitya decided to run away. Smerdyakov,

[16] In Section 2.4 I explore further the relation between *BST* and game theory.

however, did trick Fyodor into thinking that Grushenka had come, so that when Mitya did not go into the bedroom, Smerdyakov was forced to take matters into his own hands, so that it was he who killed Fyodor.

To formalize this scene using *BST*, consider the *bt*-model $\mathcal{M}$ depicted in Figure 2.3.

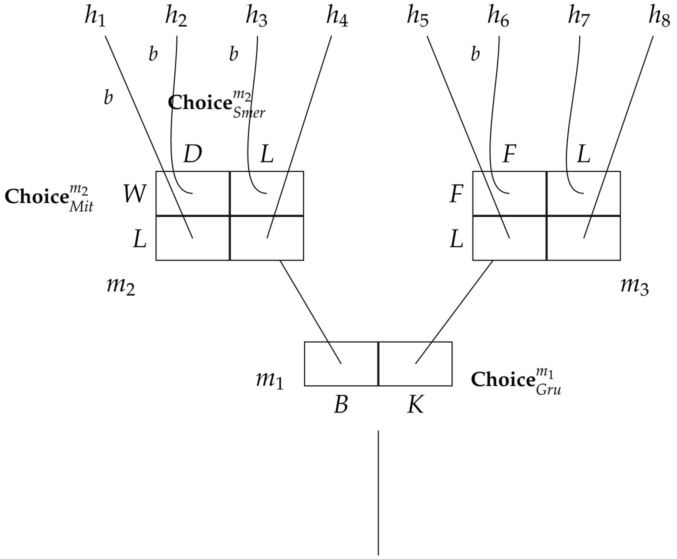


Figure 2.3: *An example of interaction.*

Here, $Ags = \{Grushenka, Mitya, Smerdyakov\}$, and m_1, m_2, and m_3 are moments, where $\sqsubset$ is defined so as to be represented by the diagram. Observe, then, that $H = \{h_i; 1 \leq i \leq 8\}$, that $H_{m_1} = H$, that $H_{m_2} = \{h_1, h_2, h_3, h_4\}$, and that $H_{m_3} = \{h_5, h_6, h_7, h_8\}$.

At m_1 *Grushenka* is faced with choosing between two actions: going to a ball (B) or going to the Karamazov household (K). Thus, $\mathbf{Choice}^{m_1}_{Gru} = \{B, K\}$. According to *Grushenka*'s choice, the world evolves toward either m_1 or m_2. Since *Mitya* and *Smerdyakov* cannot actually choose anything at m_1, their available actions are taken to lie within the trivial partition of H_{m_1}: $\mathbf{Choice}^{m_1}_{Mit} = \mathbf{Choice}^{m_1}_{Smer} = \{H_{m_1}\}$.

Both m_2 and m_3 include interdependent decision contexts involving *Mitya* and *Smerdyakov*. Intuitively, m_1 and m_2 occur at a same chronological instant—in the future of m_1—but they represent different alternatives in the possible-futures dimension of branching time.[17] As mentioned before, the layouts of choice-

[17] The idea of moments lying in same chronological instants can be formally factored into *bt*-frames. The reader is referred to Xu (2015) for details.

partitions at m_2 and m_3 look like normal-form games. Therefore, *Mitya* can be seen as the row player, and *Smerdyakov* as the column player. Since *Grushenka* cannot actually choose anything at m_2 and m_3, $\mathbf{Choice}^{m_i}_{Gru} = \{H_{m_i}\}$ for all $i \in \{2,3\}$.

At m_2 the choices available to *Mitya* are either breaking into his father's bedroom through the window (W) or leaving (L), and the choices available to *Smerdyakov* are either knocking on Fyodor's door (D) or leaving (L). Thus, $\mathbf{Choice}^{m_2}_{Mit} = \{W, L\}$ and $\mathbf{Choice}^{m_2}_{Smer} = \{D, L\}$. At m_3 the choices available to *Mitya* are either fighting his father (F) or leaving (L), and the choices available to *Smerdyakov* are also either fighting Fyodor (F) or leaving (L). Thus, $\mathbf{Choice}^{m_3}_{Mit} = \{F, L\}$ and $\mathbf{Choice}^{m_3}_{Smer} = \{F, L\}$.

Let b denote the atomic proposition 'Fyodor Karamazov is badly hurt' in Figure 2.3. The diagram then shows that

$$\mathcal{V}(b) = \{\langle m_2, h_1\rangle, \langle m_2, h_2\rangle, \langle m_2, h_3\rangle, \langle m_3, h_6\rangle\}.$$

Thus, the indices representing what actually happened in *Brothers Karamazov* are $\langle m_1, h_1\rangle$ and $\langle m_2, h_1\rangle$. At $\langle m_2, h_1\rangle$ *Mitya* has refrained from breaking into Fyodor's bedroom and *Smerdyakov* has knocked the secret code, ultimately leading to Fyodor's death. Although intuitively there is still a long sequence of actions that should occur between the moment at which Fyodor hears the code of door knocks and the moment when he gets badly hurt (namely *Smerdyakov* going into the bedroom and knocking Fyodor down with a paper-weight) in this abstraction $\mathcal{M}, \langle m_2, h_1\rangle \models b$.

Besides the actual situation, I wanted to add alternatives in the possible-futures dimension of branching time. As such, the indices anchored by h_5–h_8 concern what would have happened if that night *Grushenka* had gone to the Karamazov residence. A likely scenario is that Fyodor would not have locked himself up in his bedroom. Therefore, in Figure 2.3 the actions available to *Mitya* at m_3 are either fighting Fyodor for *Grushenka*'s love (F) or leaving (L), and the actions available to *Smerdyakov* are also either fighting Fyodor (F) or leaving (L). My reasoning for the possible outcomes at m_3 is that, since *Grushenka* was watching, Fyodor would have fought back hard. Thus, only if both *Mitya* and *Smerdyakov* had chosen F would have Fyodor been badly hurt: $\mathcal{M}, \langle m_3, h_i\rangle \models \neg b$ for all $i \in \{5, 7, 8\}$.

Now, let me exemplify the evaluation of formulas concerning agency. Observe, on the one hand, that $\mathcal{M}, \langle m_2, h_1\rangle \models \neg[Mitya]b \wedge [Smerdyakov]b$: *at the actual index it was not* Mitya *but* Smerdyakov *who hurt Fyodor*. Thus, according to our discussion on refraining (p. 38), at the actual index *Mitya* has von-Wright-refrained from seeing to it that Fyodor gets hurt, because he did not hurt Fyodor and was (causally) able to do so: $\mathcal{M}, \langle m_2, h_1\rangle \models \neg[Mitya]b \wedge \Diamond[Mitya]b$. On the other hand, observe that $\mathcal{M}, \langle m_2, h_i \rangle \models \neg\Diamond[Smerdyakov]\neg b \wedge \neg\Diamond[Mitya]\neg b$ for all $1 \leq i \leq 4$: *at all indices based on* m_2 *neither* Mitya *nor* Smerdyakov *was causally able to prevent that Fyodor got hurt.* Other interesting cases appear in the non-actual situations given by H_{m_3}. For instance, observe that $\mathcal{M}, \langle m_3, h_6 \rangle \models b \wedge \neg[Mitya]b \wedge \neg[Smerdyakov]b$: *at* $\langle m_3, h_6 \rangle$ *Fyodor has been badly hurt, but neither* Mitya *nor* Smerdyakov *did independently see to it that Fyodor got hurt.* Furthermore, $\mathcal{M}, \langle m_3, h_i \rangle \models \neg\Diamond[Mitya]b \wedge \neg\Diamond[Smerdyakov]b$ for all $5 \leq i \leq 8$: *at all indices based on* m_3 *neither* Mitya *nor* Smerdyakov *was causally able to hurt Fyodor*. Thus, although there is no explicit modelling of collaboration, one could in principle say that at $\langle m_3, h_6 \rangle$ *Mitya* and *Smerdyakov* collectively, but not individually, hurt Fyodor (see Duijf, 2018, Chapters 4 & 5, for discussions on joint collaboration and the relation between individual and collective responsibility).

This example highlights one of the benefits of stit theory that was mentioned in Chapter 1 (p. 10): the flexibility and specificity of the models. Although some might say that formalizing a scene of *Brothers Karamazov* as a case of sequential and concurrent decision-making can be a bit of a stretch, *bt*-models allow us to model a wide variety of scenarios.[18] Moreover, this small example also bears a nice connection with the discussions on responsibility that I will engage in in succeeding chapters. For instance, one could say that at $\langle m_2, h_1 \rangle$ *Smerdyakov* is causally responsible for the damage done to Fyodor.[19]

2.2.5 Deliberative Agency

The operator $[\alpha]$—with the semantics for formulas of the form $[\alpha]\varphi$ stated in Definition 2.3—is known as the Chellas-stit operator. The reason for this name is that $[\alpha]$ works as an analog of an operator that Chellas (1969) introduced to formalize agency (see Horty, 2001; Horty & Belnap, 1995). Due to its importance in the context of the formal theory of responsibility that I develop in Chapter 6, it will prove helpful to introduce a well-known variant of the Chellas-stit operator, known as the *deliberative-stit operator*:

Definition 2.6 (Deliberative-stit operator)**.** *Denoted by* $[\alpha]^d$*, the semantics for formulas involving the* deliberative-stit operator *are obtained by extending the recursive definition in Definition 2.3 with the following clause:*

$$\begin{array}{lll} \mathcal{M}, \langle m, h \rangle \models [\alpha]^d\varphi & \textit{iff} & \textit{for all } h' \in \mathbf{Choice}^m_\alpha(h), \mathcal{M}, \langle m, h' \rangle \models \varphi, \\ & & \textit{and there is } h'' \in H_m \textit{ s. t. } \mathcal{M}, \langle m, h'' \rangle \not\models \varphi. \end{array}$$

[18] In fact, the idea behind using such a scene is that any course of events that involves multi-agent choice can be modelled with stit theory. In the modelling endeavor, then, the modeller can be either as idealistic or as realistic as they please.

[19] See Chapter 1's Section 1.1 for an informal definition of *causal responsibility*. A formal account of this kind of responsibility is presented in Chapter 6's Section 6.3.

The idea is that at a given index agent α has brought about φ—by performing some action—only if φ was not already settled. Otherwise, state of affairs φ did not really depend on α's choice and therefore should not be thought of as an effect of any such choice. For $[\alpha]^d\varphi$ to hold at some index, then, there are two requirements: (1) φ must be an effect of the choice that α has performed at the index, known as the *positive condition*; and (2) $\neg\varphi$ must be historically possible, known as the *negative condition*. According to (Horty, 2001, Chapter 2), the conjunction of these two conditions is what characterizes a deliberative choice, where the term 'deliberative' comes from Thomason's (1981) notion of deliberative obligation and goes back to Aristotle's observation in *Nicomachean Ethics* (1139b7) that one can properly be said to deliberate only about what is "capable of being otherwise" (see Crisp, 2014). The deliberative-stit operator was first introduced by von Kutschera (1986) and independently by Horty (1989). For a comprehensive review of its features, the reader is referred to Horty and Belnap (1995).

Importantly, Xu (2015) pointed out that one should not confuse *deliberative action* with *deliberate action*. Modality $[\alpha]^d\varphi$ does not mean to express that α has deliberately seen to it that φ, since such a claim is intuitively related to conscious or intentional choices. Rather than expressing any of these, $[\alpha]^d\varphi$ captures the idea that φ can be seen as a true consequence of some choice available to α, whether conscious or unconscious, or intentional or unintentional.[20]

To illustrate how $[\alpha]^d$ works, consider the *bt*-model $\mathcal{M}$ depicted in Figure 2.3. Observe that $\mathcal{M}, \langle m_2, h_1\rangle \models [Smerdyakov]^d b$: *at* $\langle m_2, h_1\rangle$ Smerdyakov *has deliberatively seen to it the Fyodor is badly hurt*. The positive condition is given by the choice of D, and the negative condition is witnessed by h_4: since $\mathcal{M}, \langle m_2, h_4\rangle \models \neg b$, then $\mathcal{M}, \langle m_2, h_1\rangle \models \Diamond\neg b$: *at* $\langle m_2, h_1\rangle$ *it was not settled that Fyodor would get hurt*.

Now, there is a specific relation between the Chellas-stit operator and the deliberative-stit operator. From Definition 2.6, one can see that $[\alpha]^d\varphi$ holds at some index iff $[\alpha]\varphi$ and $\Diamond\neg\varphi$ hold at the index.[21] Indeed, these operators are interdefinable: using a language with only Chellas-stit operators, one can set $[\alpha]^d\varphi := [\alpha]\varphi \wedge \Diamond\neg\varphi$; and using a language with only deliberative-stit operators, one can set $[\alpha]\varphi := [\alpha]^d\varphi \vee \Box\varphi$.[22]

[20] Of course, these considerations are extremely significant in responsibility attribution (see, for instance, Lorini et al., 2014). To clarify, a fair question to ask is whether an agent should be held responsible for a state of affairs that was already settled. The intuitive answer is 'no,' and this implies that the formal account of deliberative action will play an important role in Chapter 6's formalization of responsibility.

[21] This means that if one extends $\mathcal{L}$ so that formulas of the form $[\alpha]^d\varphi$ are also included then formula $[\alpha]^d\varphi \leftrightarrow [\alpha]\varphi \wedge \Diamond\neg\varphi$ would be valid with respect to the class of *bt*-models.

[22] Horty (2001, Chapter 2) characterized *von Wright's kind of refraining* in terms of the deliberative-stit operator, namely through formula $\neg[\alpha]^d\varphi \wedge \Diamond[\alpha]^d\varphi$. In light of the equivalence given by $[\alpha]^d\varphi :=$

2.3 Logic-Based and Metalogic Properties of *BST*

I have overemphasized that an essential aim of this thesis's methodology is to reason about a philosophical concept's properties by means of valid/invalid formulas. I refer to such properties as *logic-based properties*. Furthermore, a very important aspect of said methodology is to develop sound and complete proof systems (otherwise known as axiomatizations) for the logics that formalize the philosophical concepts studied here. This aspect refers to the so-called *metalogic properties* of said logics. In this section I discuss important logic-based and metalogic properties of *BST*.

From here on, I will use the term *atemporal* $\mathcal{L}$ to refer to the fragment of $\mathcal{L}$ that does not include the temporal modalities $\mathtt{G}\varphi$ and $\mathtt{H}\varphi$. Thus, the name *atemporal BST* will be used to refer to the logic that is obtained by evaluating the formulas of atemporal $\mathcal{L}$ over *bt*-models, and the name *temporal BST* will be used to refer to the logic whose language does include $\mathtt{G}\varphi$ and $\mathtt{H}\varphi$.

An outline of this section should help the reader identify its purposes and contents:

- Subsection 2.3.1 explores logic-based and metalogic properties of atemporal *BST*.
- Subsection 2.3.2 discusses *Kripke semantics for atemporal* BST, comparing this semantics' ontology of agency with that of other well-known formalisms in branching-time logic.
- Subsection 2.3.3 introduces a proof system for temporal *BST* and addresses its soundness & completeness results (with respect to Kripke models). To provide a background for topics addressed later on in the thesis, a special temporal stit-theoretic formalism known as *xstit theory* is also reviewed.

2.3.1 Proof Systems for Atemporal *BST*

The experienced reader will notice that $[\alpha]$ is an **S5** modal operator for every $\alpha \in Ags$. This means that the following formulas—known as the **S5** schemata—

$[\alpha]\varphi \wedge \Diamond\neg\varphi$, $\neg[\alpha]^d\varphi \wedge \Diamond[\alpha]^d\varphi$ translates into a formula that is logically equivalent to $\neg[\alpha]\varphi \wedge \Diamond[\alpha]\varphi$. The latter formula is precisely the one that was presented to characterize von-Wright-refraining in the present chapter (p. 39).

are valid with respect to the class of *bt*-models:

$$\begin{array}{ll}
[\alpha](\varphi \to \psi) \to ([\alpha]\varphi \to [\alpha]\psi) & (K) \\
[\alpha]\varphi \to \varphi & (T) \\
[\alpha]\varphi \to [\alpha][\alpha]\varphi & (4) \\
\neg[\alpha]\varphi \to [\alpha]\neg[\alpha]\varphi & (5).
\end{array}$$

It is a matter of routine to show that, for each $\alpha \in Ags$, these formulas are valid with respect to the class of *bt*-models.[23] The reason is that any modal operator based on an equivalence relation on a set of indices implies their validity. Let me discuss what the validity of the **S5** axioms for $[\alpha]$ says about *BST*'s agency:

- The validity of (K) implies that at an index an agent will always have brought about all the logical consequences of the effects of its choice of action.

- The validity of (T) means that if at an index an agent has seen to it that φ then φ must be true at the index.

- The validity of (4) means that if at an index an agent has seen to it that φ then all the histories within the agent's current choice anchor indices at which the agent has also seen to it that φ. In other words, to have seen to it that φ is an effect of having seen to it that φ.

- The validity of (5) means that if at an index an agent has refrained from seeing to it that φ then such a refraining was also an effect of the agent's current choice.

As for the historical-necessity operator $\Box$, it is easy to see that it is also an **S5** operator.[24] The philosophical implications of the validity of the **S5** axioms for $\Box$, then, can be summarized as follows: since (K) is valid, all the logical consequences of whatever was settled at an index were also settled at that index. The validity of (T) implies that if φ was settled at an index then φ must hold at any index based on the same moment. The validity of (4) implies that if φ was settled at an index then φ's being settled was also settled at that index. Finally, the validity of (5) implies that if φ was historically possible at an index then this historical possibility was settled at the index.

[23] When the **S5** schemata are valid for a modal operator, one often refers to the operator as an **S5** operator.

[24] Furthermore, one can see $\Box$ as being based on an equivalence relation. For *bt*-model $\mathcal{M}$, let $R_\Box$ be a relation defined on $I(M \times H)$ by the following rule: $\langle m,h\rangle\, R_\Box\, \langle m',h'\rangle$ iff $m' = m$. Let us define a modality $\Delta\varphi$ as follows: $\mathcal{M}, \langle m,h\rangle \models \Delta\varphi$ iff $\mathcal{M}, \langle m',h'\rangle \models \varphi$ for every $\langle m',h'\rangle$ such that $\langle m,h\rangle\, R_\Box\, \langle m',h'\rangle$. Then $\mathcal{M}, \langle m,h\rangle \models \Delta\varphi$ iff $\mathcal{M}, \langle m,h\rangle \models \Box\varphi$ for every index $\langle m,h\rangle$.

Much more interestingly, atemporal *BST* includes principles governing the interplay between agency and historical necessity. These principles are given as non-trivial formulas that involve both modalities and that are valid with respect to *bt*-models. For instance, the semantics of $\Box\varphi$ and $[\alpha]\varphi$ found in Definition 2.3 imply that $\Box\varphi \to [\alpha]\varphi$ is valid for every $\alpha \in Ags$. Thus, whatever was settled at an index has been brought about by every agent at said index. I refer to the family of formulas of the form $\Box\varphi \to [\alpha]\varphi$ (where α ranges over Ags) as (*SET*).

Doubtlessly, the most important principle governing the interplay between $[\alpha]$ and $\Box$ is given by a family of formulas that I refer to as (*IA*). The reason for this label is that the validity of this family (with respect to *bt*-models) is closely related to frame condition *independence of agency*. This family of formulas (*IA*) is defined as follows:

$$\text{For all } m \geq 1 \text{ and pairwise distinct } \alpha_1, \ldots, \alpha_m \in Ags,$$
$$\bigwedge_{1 \leq i \leq m} \Diamond[\alpha_i]\varphi_i \to \Diamond\left(\bigwedge_{1 \leq i \leq m} [\alpha_i]\varphi_i\right) \qquad (IA)$$

One can easily show that (*IA*) is valid with respect to *bt*-models. As mentioned above, the argument involves *independence of agency*. It can also be shown that (*IA*) *defines* this frame condition, since (*IA*) is valid on a frame iff the frame satisfies *independence of agency* (see Blackburn et al., 2002, Chapter 3, for the precise definitions of frame definability through modal formulas). Therefore, one can say that (*IA*) characterizes syntactically that for each agent the availability of choices at a given moment does not depend on the choices of other agents.

As for metalogic properties, there is a Hilbert-style proof system for atemporal *BST* that is sound and complete with respect to *bt*-models. This proof system was first presented by Xu (1994), who showed it to be sound and complete, as well as decidable.[25] I define a pertinent variation of such a system below, in terms of the operators that have been introduced so far.

Definition 2.7 (Proof system for atemporal *BST*)**.** *Let Λ be the proof system defined by the following axioms and rules of inference:*

- (Axioms)
  - *All classical tautologies from propositional logic.*
  - *The* **S5** *schemata for $\Box$ and $[\alpha]$ (for each $\alpha \in Ags$).*

[25] On the basis of Wölfl's (2002) axiomatization of temporal *BST* (where the language includes formulas involving two extra modal operators besides $\Box$, $[\alpha]$, $\mathtt{G}$, and $\mathtt{H}$), one can produce an alternative axiomatization for atemporal *BST*. Two further axiomatizations for atemporal *BST*, equivalent to the one proposed by Xu (1994), were given by Balbiani, Herzig, and Troquard (2008). Additionally, a complete tableaux calculus for this logic was provided by Wansing (2006b).

- *The family of schemata* (*SET*).
- *The family of schemata* (*IA*).

- (Rules of inference) *Modus Ponens, Substitution, and Necessitation for the modal operators.*

The main metalogic result for Λ, then, is given by the following theorem:

Theorem 2.8 (Soundness & Completeness of Λ, Xu, 1994)**.** *The proof system Λ is sound and complete with respect to the class of bt-models. Furthermore, it is decidable.*

In Xu's proof of soundness & completeness, the spotlight falls on axiom (*IA*). Since (*IA*) is valid with respect to *bt*-models, soundness of Λ is a straightforward result. As for the proof of completeness, it will prove useful to review Xu' strategy, since it is similar to those underlying all the proofs of completeness in this thesis. Xu' strategy involves the following steps:

- (Step 1) A *canonical frame* $\mathcal{F}$ for Λ is built just as in modal logic. Therefore, this canonical frame is a Kripke structure based on a domain of possible worlds. Each possible world is a Λ-MCS, where 'MCS' stands for 'maximally consistent set of formulas.' Equivalence relations $R_\Box$ and R_α (for each $\alpha \in Ags$) are defined on said domain so that, for world w, $\Box\varphi \in w$ iff $\varphi \in v$ for every v such that $wR_\Box v$, and, for $\alpha \in Ags$, $[\alpha]\varphi \in w$ iff $\varphi \in v$ for every v such that $wR_\alpha v$.

- (Step 2) For a world w in the canonical frame, a *bt*-model $\mathcal{M}_w$ is built as follows: the domain of $\mathcal{M}_w$ includes a root moment $m_w = \{w\}$ and all the worlds v that lie in the $R_\Box$-class of w. Each v in the $R_\Box$-class of w corresponds to an index of the form $\langle m_w, h_v\rangle$ (where history $h_v = \{m_w, v\}$). Function **Choice** is then defined on the basis of the R_α's. This definition, coupled with the fact that $(IA) \in v$ for every world v, ensures that frame condition *independence of agency* is satisfied in $\mathcal{M}_w$. Therefore, $\mathcal{M}_w$ is a *bt*-model in all the sense of the word. The valuation function $\mathcal{V}$ for $\mathcal{M}_w$ is defined so that $\mathcal{M}, \langle m_w, h_v\rangle \models \varphi$ iff $\varphi \in v$. For each Λ-consistent formula φ, Lindenbaum's Lemma (Blackburn et al., 2002, Chapter 4, p. 199) implies that φ is included in some w' in the domain of the canonical frame. Thus, for each Λ-consistent formula φ there exists a *bt*-model $\mathcal{M}_{w'}$ such that $\mathcal{M}_{w'}, \langle m_{w'}, h_{w'}\rangle \models \varphi$. This means that Λ is complete with respect to the class of *bt*-models.

As for Xu's proof of decidability, it is based on the *finite model property*. The existence of a finite model is shown with a standard argument in modal logic: *filtration* (see Blackburn et al., 2002, Chapters 2, 4, 6, for an exposition of finitary methods, filtrations, and the decidability of a variety of modal logics).

As is often the case, interesting valid formulas result from soundness of a proof system. For instance, an important consequence of the validity of (*SET*), (*IA*), and the **S5** schemata for $\Box$ and $[\alpha]$ is that the modality for historical necessity can be expressed in terms of the modalities for agency, provided that *Ags* includes more than one agent. In other words, the validity of the mentioned schemata implies that, for all $\alpha, \beta \in Ags$ such that $\alpha \neq \beta$, $\Box\varphi \leftrightarrow [\alpha][\beta]\varphi$ is valid. This formula says that an agent has seen to it that other agent brings about φ iff φ was already settled. To put it colloquially, an agent cannot have forced another agent to bring about any state of affairs unless such a state was already settled. To illustrate the deduction of Λ-theorems, a derivation of $\Box\varphi \leftrightarrow [\alpha][\beta]\varphi$ is included below, where 'Subs.' abbreviates 'Substitution.'[26]

1. $\vdash_\Lambda \Box[\beta]\varphi \leftrightarrow [\beta]\Box\varphi$	(4) for $\Box$, subs. of (SET), subs. of (T) for $\Box$, subs. of (T) for $[\beta]$
2. $\vdash_\Lambda \Diamond\langle\beta\rangle\varphi \leftrightarrow \langle\beta\rangle\Diamond\varphi$	Contrapositive of 1
3. $\vdash_\Lambda \Diamond\varphi \rightarrow \langle\beta\rangle\Diamond\varphi$	Subs. of (T) for $[\beta]$
4. $\vdash_\Lambda \Diamond\varphi \rightarrow \Diamond\langle\beta\rangle\varphi$	3, 2, prop. logic
5. $\vdash_\Lambda \Diamond\varphi \wedge [\alpha][\beta]\neg\varphi \rightarrow \Diamond\langle\beta\rangle\varphi \wedge [\alpha][\beta]\neg\varphi$	4, prop. logic
6. $\vdash_\Lambda \Diamond\langle\beta\rangle\varphi \wedge [\alpha][\beta]\neg\varphi \rightarrow \Diamond\langle\beta\rangle\varphi \wedge \Diamond[\alpha][\beta]\neg\varphi$	Subs. of (T) for $\Box$, prop. logic
7. $\vdash_\Lambda \Diamond\langle\beta\rangle\varphi \wedge \Diamond[\alpha][\beta]\neg\varphi \rightarrow \Diamond[\beta]\langle\beta\rangle\varphi \wedge \Diamond[\alpha][\beta]\neg\varphi$	Subs. of (5) for $[\beta]$, modal logic
8. $\vdash_\Lambda \Diamond[\beta]\langle\beta\rangle\varphi \wedge \Diamond[\alpha][\beta]\neg\varphi \rightarrow \Diamond([\beta]\langle\beta\rangle\varphi \wedge [\alpha][\beta]\neg\varphi)$	Subs. of (IA)
9. $\vdash_\Lambda \Diamond([\beta]\langle\beta\rangle\varphi \wedge [\alpha][\beta]\neg\varphi) \rightarrow \Diamond(\langle\beta\rangle\varphi \wedge [\beta]\neg\varphi)$	Subs. of (T) for $[\beta]$, Subs. of (T) for $[\alpha]$, modal logic
10. $\vdash_\Lambda \Diamond(\langle\beta\rangle\varphi \wedge [\beta]\neg\varphi) \rightarrow \Diamond\bot$	Modal logic
11. $\vdash_\Lambda \Diamond\varphi \wedge [\alpha][\beta]\neg\varphi \rightarrow \bot$	5–10, prop. logic
12. $\vdash_\Lambda \Diamond\varphi \rightarrow \langle\alpha\rangle\langle\beta\rangle\varphi$	11, prop. logic
13. $\vdash_\Lambda [\alpha][\beta]\varphi \rightarrow \Box\varphi$	Contrapositive of 12
14. $\vdash_\Lambda \Box\varphi \rightarrow [\alpha][\beta]\varphi$	(4) for $\Box$, subs. of (SET)
15. $\vdash_\Lambda \Box\varphi \leftrightarrow [\alpha][\beta]\varphi$	13, 14, prop. logic

[26] Of course, one can show that formula $\Box\varphi \leftrightarrow [\alpha][\beta]\varphi$ is valid using only a semantic argument. Unsurprisingly, the proof relies on *independence of agency*. Completeness of Λ with respect to *bt*-models then yields that such a formula is a Λ-*theorem*.

2.3.2 Kripke Semantics for Atemporal *BST*

In Xu's (1994) proof of completeness, the first step points to the possibility of giving an alternative semantics to the formulas of atemporal $\mathcal{L}$, namely in terms of relations on Kripke structures of possible worlds. Indeed, such a semantics, which I will henceforward refer to as *Kripke semantics for atemporal* BST, has been extensively explored in the literature, for a variety of reasons. Since it plays an important role in this thesis, let me address the basic ideas behind it.

First formally introduced by Balbiani et al. (2008), and independently by Kooi and Tamminga (2008), Kripke semantics for atemporal *BST* has often been used to work with simpler models (see, for instance, Broersen, 2011a; Duijf, 2018; Herzig & Schwarzentruber, 2008; Schwarzentruber, 2012). Indeed, one can express all the logic-based properties of atemporal *BST* in terms of Kripke semantics, since—as discussed below—the proof system Λ (Definition 2.7) is sound and complete with respect to certain classes of Kripke structures.

Although the works mentioned in the previous paragraph introduced Kripke semantics for atemporal *BST* in slightly different ways, they all presuppose domains of possible worlds and accessibility relations on them. The following definition of *Kripke*-stit *frames* closely follows the presentation of Balbiani et al. (2008) (see also Broersen, 2011a; Duijf, 2018).

Definition 2.9 (Kripke-*stit*-frames & models). *A tuple* $\langle W, Ags, R_\Box, \mathtt{Choice}\rangle$ *is called a* Kripke-*stit*-frame *iff*

- *W is a non-empty set of possible worlds. $R_\Box$ is an equivalence relation over W. For $w \in W$, the class of w under $R_\Box$ is denoted by $\overline{w}$.*
- $\mathtt{Choice}$ *is a function that assigns to each $\alpha \in Ags$ and $\Box$-class $\overline{w}$ a partition* $\mathtt{Choice}^{\overline{w}}_\alpha$ *of $\overline{w}$ given by an equivalence relation denoted by $R^{\overline{w}}_\alpha$.* $\mathtt{Choice}$ *must satisfy the following constraint:*
  - $(\mathtt{IA})_{\mathrm{K}}$ *For all $w \in W$, each function $s : Ags \to 2^{\overline{w}}$ that maps α to a member of* $\mathtt{Choice}^{\overline{w}}_\alpha$ *is such that $\bigcap_{\alpha \in Ags} s(\alpha) \neq \emptyset$ (where, mirroring Definition 2.2, the set of all functions s that map α to a member of* $\mathtt{Choice}^{\overline{w}}_\alpha$ *is denoted by* $\mathtt{Select}^{\overline{w}}$*).*

For $\alpha \in Ags$, $w \in W$, and $v \in \overline{w}$, the class of v in the partition $\mathtt{Choice}^{\overline{w}}_\alpha$ *is denoted by* $\mathtt{Choice}^{\overline{w}}_\alpha(v)$*.*

A Kripke-*stit*-model $\mathcal{M}$*, then, is a tuple that results from adding a valuation function $\mathcal{V}$ to a Kripke-*stit*-frame, where $\mathcal{V} : P \to 2^W$ assigns to each atomic proposition a set of possible worlds (recall that P is the set of propositions in atemporal $\mathcal{L}$).*

Kripke-*stit*-models are used to evaluate the formulas of $\mathcal{L}$ with semantics that are analogous to those provided using *bt*-models:

Definition 2.10 (Evaluation rules on Kripke models)**.** *Let $\mathcal{M}$ be a Kripke-*stit*-model. The semantics on $\mathcal{M}$ for the formulas of $\mathcal{L}$ are defined recursively by the following truth conditions, evaluated at world w:*

$$\begin{array}{lll}
\mathcal{M}, w \models p & \textit{iff} & w \in \mathcal{V}(p) \\
\mathcal{M}, w \models \neg\varphi & \textit{iff} & \mathcal{M}, w \not\models \varphi \\
\mathcal{M}, w \models \varphi \wedge \psi & \textit{iff} & \mathcal{M}, w \models \varphi \textit{ and } \mathcal{M}, w \models \psi \\
\mathcal{M}, w \models \Box\varphi & \textit{iff} & \textit{for all } v \in \overline{w}, \mathcal{M}, v \models \varphi \\
\mathcal{M}, w \models [\alpha]\varphi & \textit{iff} & \textit{for all } v \in \mathtt{Choice}^{\overline{w}}_{\alpha}(w), \mathcal{M}, v \models \varphi.
\end{array}$$

Satisfiability, validity, and general validity are defined as usual.

As reviewed by Duijf (2018, Chapter 1), these Kripke-*stit*-models are very similar to (a) Kooi and Tamminga's (2008) *choice structures* , to (b) van Benthem and Pacuit's (2014) *STIT choice structures*, and to (c) Ciuni and Horty's (2014) *choice Kripke models*.

Kripke-*stit*-models are most commonly used to simplify—and also clarify—the logic-based and metalogic properties of (temporal and atemporal) *BST*. As an example, all the proofs of completeness in this thesis make heavy use of these models, following the seminal works that presented completeness results for *BST* (Balbiani et al., 2008; Broersen, 2011a; Herzig & Schwarzentruber, 2008; Lorini, 2013; Schwarzentruber, 2012; Xu, 1994). In the case of atemporal *BST*, the fact that one can 'safely' use Kripke-*stit*-models to explore the logic comes from the important proposition below.

Theorem 2.11 (Soundness & Completeness of Λ (w. r. t. Kripke models))**.** *The proof system Λ of Definition 2.7 is sound and complete with respect to the class of Kripke-*stit*-models.*

The proof of this result is well-known. Soundness is routine, and completeness follows from ordinary techniques of modal logic.[27] In turn, Theorem 2.11 implies the following standard results:

Proposition 2.12. *A formula φ of atemporal $\mathcal{L}$ is valid with respect to the class of* bt-*models iff φ is valid with respect to the class of Kripke-*stit*-models.*

[27] In fact, these techniques appear in most of the works that I have mentioned above. They refer to *canonical models* and *Sahlqvist correspondence results* (see Blackburn et al., 2002, Chapter 4, for an exposition of these techniques).

Proposition 2.13. *A formula φ of atemporal $\mathcal{L}$ is satisfiable with respect to the class of* bt-*models iff φ is satisfiable with respect to the class of Kripke-*stit-*models.*

More interestingly, certain nuances in the relation between *bt*-models, on the one hand, and Kripke-*stit* models, on the other, can be seen in Herzig and Schwarzentruber's (2008) proof of Proposition 2.13. In this proof, it is shown that (a) for a formula φ of atemporal $\mathcal{L}$ that is satisfied at an index of a given *bt*-model, one can use this *bt*-model and said index to construct an associated Kripke-*stit*-model such that φ is satisfied at a corresponding world within it; and (b) (the other way round) for a formula φ of atemporal $\mathcal{L}$ that is satisfied at a world of a given Kripke-*stit*-model, one can construct an associated *bt*-model such that φ is satisfied at a corresponding index within it. For (a), the histories passing through the pertinent index of the *bt*-model are seen as worlds of the associated Kripke-*stit*-model. For (b), the elements in the $R_\Box$-class of the pertinent world are seen as terminal nodes of histories passing through the corresponding index, such that each world in said $R_\Box$-class corresponds to a unique history.[28]

One may gather a specific insight, concerning the ontology of Kripke semantics for atemporal *BST*, from the discussion above. In the Kripke-*stit*-models of Definition 2.9, each possible world can be identified with a history passing through a moment. Therefore, if one is to adapt the conceptual ideas of agency—in terms of partitions of the set of histories passing through a moment—to Kripke-*stit*-models, then the actions available to agents would involve partitions of the set of possible worlds. As explained by Duijf (2018), this implies that each possible world should be taken to include a full temporal evolution of the world—even in this setting without temporal operators.

Let me briefly elaborate on this matter. Even without temporal operators, atemporal *BST*'s notion of agency depends on branching time. As explained by Zanardo (1996), there is an important ontological difference between modelling branching time as a set of *moments* and modelling it as a set of *possible worlds*. Although in Kripke semantics the formulas of atemporal $\mathcal{L}$ are evaluated at possible worlds, each possible world should be seen as a complex entity that includes a representation of a full course of events.[29] Only in this way can the

[28] In Chapter 4, a similar—and a bit stronger—result is shown in the context of an extension of atemporal *BST* both with epistemic and deontic operators. Definition C.43 and Propositions C.44 and C.46 imply that every Kripke-*stit*-model can be used to construct an associated *bt*-model such that there exists a bijective correspondence between the domain of the Kripke-*stit*-model and the set of histories in the associated *bt*-model, where φ is satisfied at a world in the Kripke-*stit*-model iff φ is satisfied at an index anchored by the history corresponding to said world in the associated *bt*-model.

[29] The *worlds* in *Kamp frames* for branching-time logic (Thomason, 1984; Zanardo, 1996, 2006), for instance, are based on this ontological premise. Each world, then, is associated with a complete flow of time. I will briefly return to this discussion when addressing Kripke semantics for temporal *BST*, in the next subsection.

concept of agency successfully carry over to Kripke structures. Indeed, this ontological distinction led Duijf to using the term *dynamic world* to refer to the elements in the domains of Kripke-*stit*-models. Such a term "highlights that we can speak of the action preformed by an agent at a dynamic world, whereas this would be fallacious or elusive for standard possible worlds" (Duijf, 2018, Chapter 1, p. 34).

2.3.3 Proof Systems for Temporal *BST*

The logic-based and metalogic properties of temporal *BST* (whose language includes modalities $\Box\varphi$, $[\alpha]\varphi$, $\mathtt{G}\varphi$, and $\mathtt{H}\varphi$) have also been explored in the literature (see, for instance, Ciuni & Lorini, 2017; Lorini, 2013; Wölfl, 2002). To discuss them, I present a logic that is based on Lorini's (2013) framework, where the semantics are given on special Kripke structures that I refer to as *Kripke*-tstit-*frames*.[30]

Definition 2.14 (Kripke-*tstit*-frames & models)**.** *A tuple* $\langle W, Ags, R_\Box, R_\mathtt{G}, R_\mathtt{H}, \mathtt{Choice}\rangle$ *is called a* Kripke-*tstit*-frame *iff*

- $\langle W, Ags, R_\Box, \mathtt{Choice}\rangle$ *is a Kripke-*stit*-model (Definition 2.9).*
- $R_\mathtt{G}$ *is a relation on* W*, representing the* future relation*, with the following properties:*
  - Seriality*: for all* $w \in W$*, there is* $w' \in W$ *such that* $wR_\mathtt{G}w'$*.*
  - Transitivity*: for all* $w, x, y \in W$*, if* $wR_\mathtt{G}x$ *and* $xR_\mathtt{G}y$*, then* $wR_\mathtt{G}y$*.*
  - Linearity*: for all* $w, x, y \in W$*, if* $wR_\mathtt{G}x$ *and* $wR_\mathtt{G}y$*, then either* $xR_\mathtt{G}y$ *or* $yR_\mathtt{G}x$ *or* $x = y$*.*
  - Chronological irreflexivity*: for all* $w, x \in W$*, if* $wR_\Box x$*, then it is not the case that* $wR_\mathtt{G}x$*.*
- $R_\mathtt{H}$ *is a relation on* W*, representing the* past relation*, with the following properties:*
  - Inverse*:* $R_\mathtt{H} = R_\mathtt{G}^{-1}$*.*
  - Linearity*: for all* $w, x, y \in W$*, if* $wR_\mathtt{H}x$ *and* $wR_\mathtt{H}y$*, then either* $xR_\mathtt{H}y$ *or* $yR_\mathtt{H}x$ *or* $x = y$*.*
- *Furthermore,* $\mathtt{Choice}$*,* $R_\mathtt{G}$*, and* $R_\Box$ *are such that the following condition is satisfied:*

[30] The logic presented here is not exactly the same as Lorini's. The main difference is that, while Lorini's language includes an extra modality—$[Ags]\varphi$—expressing agency of the grand coalition of agents (meaning the full set *Ags*), the logic addressed here is obtained over the language $\mathcal{L}$ of Definition 2.1. Thus, I work by restricting Lorini's language to formulas that do not involve [*Ags*].

- $(\mathtt{NC})_{\mathtt{K}}$ No choice between undivided 'histories': *for all $\alpha \in Ags$, if $wR_{\mathtt{G}}v$ and $vR_{\Box}x$, then there exists $y \in \mathtt{Choice}_{\alpha}^{\overline{w}}(w)$ such that $yR_{\mathtt{G}}x$.*

A Kripke-*tstit*-model *$\mathcal{M}$, then, is a tuple that results from adding a valuation function $\mathcal{V}$ to a Kripke-*tstit*-frame, where $\mathcal{V} : P \to 2^W$ assigns to each atomic proposition a set of possible worlds (recall that P is the set of propositions in $\mathcal{L}$).*

Definition 2.15 (Evaluation rules for temporal *BST* on Kripke models)**.** *Let $\mathcal{M}$ be a Kripke-*tstit*-model. The semantics on $\mathcal{M}$ for the formulas of $\mathcal{L}$ is defined recursively just as in Definition 2.10, with the following additional clauses:*

$$\begin{array}{lll} \mathcal{M}, w \models \mathtt{G}\varphi & \textit{iff} & \textit{for all } v \textit{ such that } wR_{\mathtt{G}}v, \mathcal{M}, v \models \varphi \\ \mathcal{M}, w \models \mathtt{H}\varphi & \textit{iff} & \textit{for all } v \textit{ such that } wR_{\mathtt{H}}v, \mathcal{M}, v \models \varphi. \end{array}$$

Carrying on with the discussion on the ontology of Kripke semantics for *BST*, observe that Kripke-*tstit*-models explicitly account for the chronological dimension of time, with $R_{\mathtt{G}}$ and $R_{\mathtt{H}}$. $R_{\mathtt{G}}$'s properties of seriality, transitivity, linearity, and chronological irreflexivity ensure that, for each $w \in W$, $R_{\mathtt{G}}$ behaves just as linear temporal logic's future relation on the set $R_{\mathtt{G}}[w] = \{v \in W; wR_{\mathtt{G}}v\}$ (see Blackburn et al., 2002, Chapter 2). The same goes for $R_{\mathtt{H}}$ as the past relation, for which the fact that $R_{\mathtt{H}} = R_{\mathtt{G}}^{-1}$ ensures that $R_{\mathtt{H}}$ is also transitive.[31] For $w \in W$, a flow of time associated with w is given by the strict linear order $R_{\mathtt{H}}[w] \cup \{w\} \cup R_{\mathtt{G}}[w]$. *Moments*, then, are identified with the $R_{\Box}$-equivalence classes. To clarify, if the set $\{\overline{w}; w \in W\}$ is ordered by an ordering $\sqsubset$ defined by $\overline{w} \sqsubset \overline{w'}$ iff $wR_{\mathtt{G}}w'$, then $R_{\mathtt{G}}$'s properties ensure that $\sqsubset$ is a well-defined strict partial ordering, where condition $(\mathtt{NC})_{\mathtt{K}}$ implies that $\sqsubset$ has no backward branching. Therefore, $\sqsubset$ behaves just as the orderings underlying *bt*-models.[32]

[31] Notice, however, that $R_{\mathtt{H}}$ might not be serial, so that there could be a 'first' moment.

[32] Kripke-*tstit*-frames are very similar to Zanardo's (1985) *Ockhamist frames* for branching-time logic. Indeed, an Ockhamist frame is a Kripke structure of possible worlds, where a union $<$ of disjoint irreflexive linear orders on the domain represents the chronological dimension of branching time, and where an equivalence relation $\sim$ on the domain (disjoint from $<$) represents moments as classes of possible worlds such that, for a given class, the members of this class embody the starting points of possible histories passing through the moment that the class represents. Thus, Kripke-*tstit*-frames are essentially obtained by extending Ockhamist frames so that they include choice partitions.

As established by Zanardo (2006) and by Ciuni and Lorini (2017), Ockhamist frames are equivalent for branching-time logic to two kinds of structures: *Kamp frames* and *bundled trees*. As for *Kamp frames*, I mentioned in Footnote 29 that the elements in the domain of a Kamp frame are called worlds. In a Kamp frame, each world w is associated with a complete *flow of time* of the form $(T_w, <_w)$, such that $<_w$ is an irreflexive, transitive, and linear ordering on a set T_w of 'moments.' The branching dimension is given by equivalence relations $\approx_t$ on the domain, where, for world w and $t \in T_w$, $\approx_t$ is defined so that the equivalence $w \approx_t w'$ is meant to imply that the flow of time associated with w matches the flow of time associated with w' up to 'moment' t. As for *bundled trees*, these are built as follows: for a strict partial order of moments $\langle M, \sqsubset \rangle$ such that $\sqsubset$ does not admit backward branching and such that

Following Lorini's (2013) presentation, an axiomatization of temporal *BST*—with respect to Kripke-*tstit*-frames—is given below.

Definition 2.16 (Proof system for temporal *BST*). *Let Λ_T be the proof system defined by the following axioms and rules of inference:*

- (Axioms)
  - *All the axioms and schemata given in Definition 2.7 for the proof system Λ; the* **KD4** *axioms for $\mathtt{G}$; the* **K** *axiom for $\mathtt{H}$; and the following axioms and schemata:*

$$\begin{array}{ll} \varphi \to \mathtt{GP}\varphi & (\mathit{In}_{\mathtt{G,H}}) \\ \varphi \to \mathtt{HF}\varphi & (\mathit{In}_{\mathtt{H,G}}) \\ \mathtt{PF}\varphi \to (\mathtt{P}\varphi \vee \varphi \vee \mathtt{F}\varphi) & (\mathit{Lin}_{\mathtt{G}}) \\ \mathtt{FP}\varphi \to (\mathtt{P}\varphi \vee \varphi \vee \mathtt{F}\varphi) & (\mathit{Lin}_{\mathtt{H}}) \\ [\alpha]\mathtt{G}\varphi \to \mathtt{G}\Box\varphi & (\mathit{NCUH}) \end{array}$$

- (Rules of inference)
  - *All the rules of inference given in Definition 2.7 for the proof system Λ.*
  - Necessitation *for $\mathtt{G}$ and $\mathtt{H}$.*
  - *The rule of* irreflexivity*: from $(\Box\neg p \wedge \Box(\mathtt{G}p \wedge \mathtt{H}p)) \to \varphi$ infer φ, provided that p does not occur in φ.*

Therefore, Λ_T is obtained by extending Λ with axioms that syntactically characterize temporal logic's traditional properties for $R_{\mathtt{G}}$ and $R_{\mathtt{H}}$. Additionally, schema (*NCUH*) characterizes syntactically frame condition $(\mathtt{NC})_{\mathtt{K}}$ (no choice between undivided histories). The rule of irreflexivity is a variant of Gabbay's well-known irreflexivity rule (Gabbay, Hodkinson, & Reynolds, 1994), which according to Lorini (2013) has been widely used for proving completeness results for different temporal logics in which time is assumed to be irreflexive (see, for instance, Gabbay et al., 1994; von Kutschera, 1997; Zanardo, 1996).

The main metalogic result for the proof system Λ_T is given by the following theorem:

the set of histories is denoted by H (where H is defined just as in *bt*-frames), a *bundle* $\mathcal{B}$ is a subset of H such that for each $m \in M$ there is $h \in \mathcal{B}$ such that $m \in h$. For any bundle $\mathcal{B}$, the tuple $\langle M, \mathcal{B}\rangle$ is known as a *bundled tree*. Since Kamp frames and bundled trees were shown to be equivalent for branching-time logic to Ockhamist frames, one may regard Kripke-*tstit*-frames as extensions of any of these two kinds of structures with choice partitions, where these partitions refine the equivalents of $R_\Box$-equivalence classes in such structures. Indeed, Ciuni and Lorini (2017) already presented such extensions: *agent Kamp frames* and *choice b-trees*, respectively. All these comparisons, then, allow us to elucidate the temporal nature of Kripke-*tstit*-frames, for which it seems appropriate to identify worlds with histories, on the one hand, and $R_\Box$-equivalence classes with moments, on the other.

Theorem 2.17 (Soundness & Completeness of Λ_T, Lorini, 2013)**.** *The proof system Λ_T is sound and complete with respect to the class of Kripke-*tstit*-models.*

A fair question to ask is whether Λ_T is also sound and complete with respect to *bt*-models. The answer is negative. As implied by Ciuni and Lorini's (2017) correspondence results, a formula φ of $\mathcal{L}$ is valid with respect to Kripke-*tstit*-models iff φ is valid with respect to *agent Kamp frames*, which in turns happens iff φ is valid with respect to *choice b-trees* (see Footnote 32). This means that Λ_T is also sound and complete with respect to both these classes. However, a well-known fact (see, for instance, Burgess, 1979; Zanardo, 2006) is that there are formulas of $\mathcal{L}$ that are valid with respect to *bt*-models but invalid with respect to *choice b-trees*. An example of one such formula is $\Box \mathtt{G} \Diamond \mathtt{F} \Box \varphi \rightarrow \Diamond \mathtt{GF} \Box \varphi$. Thus, soundness of Λ_T with respect to *choice b-trees* implies that Λ_T is not complete with respect to *bt*-models.[33]

2.3.3.1 A Special Kind of Temporal Stit Theory: *Xstit*

Following ideas first presented by Herzig and Troquard (2006), Broersen (2008a) (see also Broersen, 2011a) introduced a special kind of temporal stit theory to reason about the relation between knowledge and agency: *Xstit*. In this logic, an agent's choices of action, rather than having instantaneous effects, affect next moments. I want to include a discussion of *Xstit* for two reasons:

1. The assumption that actions affect next moments implies that agents' choices can be seen as underlying transitions between states in a discrete temporal structure. To clarify, having seen to it that φ is equated with performing an action whose transition results in a state at which φ holds. Therefore, *Xstit* draws a bridge between stit theory and prominent logics for multi-agent systems (see Subsection 2.4.2).

2. *Xstit* is of particular relevance for Chapter 3, where the stages of information disclosure in interactive decision contexts are factored into stit theory to complement the theory of agency with epistemic notions from epistemic game theory (*ex ante*, *ex interim*, and *ex post* knowledge).

[33] It is worth mentioning that Wölfl (2002) provided an axiomatization of an analog of temporal *BST* known as *propositional q-logic*. Using two extra modal operators besides operators analogous to the ones of language $\mathcal{L}$, Wölfl showed that propositional q-logic is sound and complete with respect to *$T \times W$-based agent-frames*, which are nothing more than agent Kamp frames where all the flows of time coincide with a single linear order (Zanardo, 2006). The two extra operators respectively express that a state of affairs is settled throughout the current *instant* (see Footnote 17) and that a state of affairs holds along every history passing through the current moment except the current history (see Xu, 2015, Footnote 11, p. 855).

As for *Xstit*'s syntax, Broersen (2008a) introduced modality $[\alpha]^X\varphi$ to express that α has seen to it that φ will occur at the next moment. As for the semantics, the idea of next moments implies that in the frames used to evaluate *Xstit* formulas time is discrete. Therefore, I refer to these frames as *branching-discrete-time frames*.[34] The formal definitions are included below, for which the closest formulation in the literature is Xu's (2015).

Definition 2.18 (Syntax of *Xstit*)**.** *Given a finite set Ags of agent names and a countable set of propositions P, the grammar for the formal language $\mathcal{L}_{\mathsf{X}}$ is given by*

$$\varphi ::= p \mid \neg\varphi \mid \varphi \wedge \varphi \mid \Box\varphi \mid X\varphi \mid [\alpha]^X\varphi,$$

where p ranges over P and α ranges over Ags.

In $\mathcal{L}_{\mathsf{X}}$, $\Box\varphi$ has the same meaning as in *BST*; $X\varphi$ expresses that 'φ holds at the next moment (along the same history)'; and $[\alpha]^X\varphi$ expresses 'α has seen to it that φ will hold at the next moment, along the same history.'

Definition 2.19 (*Bdt*-frames & models)**.** *A tuple $\langle M, \sqsubset, Ags, \mathbf{Choice}\rangle$ is called a* branching-discrete-time frame *(*bdt*-frame for short) iff*

- *M is a non-empty set of moments and $\sqsubset$ is a strict partial ordering on M satisfying 'no backward branching.' In contrast to the frames of Definition 2.2, these structures are called 'discrete-time' because $(M, \sqsubset)$ must meet the following requirement:*
  - (TD) Time-discreteness*: for all $m \in M$ and $h \in H_m$, there exists a unique moment m^{+h} such that $m \sqsubset m^{+h}$ and $m^{+h} \sqsubseteq m'$ for every $m' \in h$ such that $m \sqsubset m'$. For $m \in M$ and $h \in H_m$, m^{+h} is known as the* successor of *m* along *h. For an index $\langle m, h\rangle$, I refer to $\left\langle m^{+h}, h\right\rangle$ as the* successor of $\langle m, h\rangle$.
- **Choice** *is a function defined just as in Definition 2.2.*[35]

A bdt-model $\mathcal{M}$*, then, is a tuple that results from adding a valuation function $\mathcal{V}$ to a* bdt*-frame, where $\mathcal{V} : P \to 2^{I(M\times H)}$ assigns to each atomic proposition a set of indices.*

[34] Xu (2015) referred to these frames as *incremental stit frames*.

[35] Broersen's (2008a) original formulation of *Xstit* frames is a bit different. Instead of choice-partitions, Broersen introduced, for each agent α, a serial relation R_α on $I(M \times H)$ such that, for all $m \in M$ and $h \in H$, $\langle m, h\rangle R_\alpha \langle m', h'\rangle$ implies that m' is a successor of m along a history h' that passes both through m' and m and that implicitly lies within the same action as h. Broersen also introduced a serial and deterministic 'next-time' relation R_{Ags}, under the premises that (a) R_{Ags} underlies a single transition between indices, and that (b) this transition is determined by an action of the full set of agents *Ags*, so that $R_{Ags} \subseteq R_\alpha$ for every $\alpha \in Ags$. As such, R_α is the relation underlying modality $[\alpha]^X\varphi$, and R_{Ags} is the relation underlying modality $X\varphi$. As established by Xu (2015), both formulations are logically equivalent, but in Broersen's presentation the choice-partitions only "appear (implicitly) as sets of possible *next* states" (Broersen, 2008a, p. 50, emphasis in original).

Definition 2.20 (Evaluation rules for *Xstit*). *Let $\mathcal{M}$ be a* bdt-*model. The semantics on $\mathcal{M}$ for the formulas of $\mathcal{L}_X$ are defined by extending the recursive definition in Definition 2.3 with the following clauses:*

$$\begin{array}{lll} \mathcal{M}, \langle m,h \rangle \models X\varphi & \textit{iff} & \mathcal{M}, \left\langle m^{+h}, h \right\rangle \models \varphi \\ \mathcal{M}, \langle m,h \rangle \models [\alpha]^X\varphi & \textit{iff} & \textit{for all } h' \in \mathbf{Choice}^m_\alpha(h), \mathcal{M}, \left\langle m^{+h'}, h' \right\rangle \models \varphi. \end{array}$$

There are alternative presentations of *Xstit* that treat $[\alpha]^X$ as a fused operator of the traditional $[\alpha]$ with temporal operator X (see, for instance, Herzig & Troquard, 2006; Schwarzentruber, 2012; Xu, 2015). Indeed, if $\mathcal{L}_X$ is defined by extending atemporal $\mathcal{L}$ with modality $X\varphi$, and if the semantics for this modality is just as in Definition 2.20 above, then the resulting logic (over *bdt*-models) is rich enough to express choices of action both with instantaneous effects—via $[\alpha]\varphi$—and with effects at next moments—via $[\alpha]X\varphi$.[36]

Definition 2.21 (Alternative syntax for *Xstit*). *Given a finite set Ags of agent names and a countable set of propositions P, the grammar for the formal language $\mathcal{L}_X$ is given by*

$$\varphi ::= p \mid \neg\varphi \mid \varphi \wedge \varphi \mid \Box\varphi \mid X\varphi \mid [\alpha]\varphi,$$

where p ranges over P and α ranges over Ags.

Strictly speaking, this 'alternative presentation' of *Xstit* is in fact a richer logic than Broersen's (2008a). To clarify, the latter can be embedded in the 'alternative' through a translation Tr such that $Tr\left([\alpha]^X\varphi\right) = [\alpha]X\varphi$. Therefore, I refer to the alternative presentation as *basic xstit theory*. As for metalogic results for these logics, they can be summarized as follows:

- The technical results by Broersen (2008a) and by Payette (2014) imply that there is a proof system for *Xstit* that is sound and complete with respect to a class of Kripke models (as well as decidable). These models are extensions of Kripke-*stit*-models such that each world in the domain has a next world under a serial and deterministic relation R_X.

- Schwarzentruber's (2012) results imply that there is a proof system for *basic xstit theory* that is sound and complete with respect to *bdt*-models, as well as decidable.

[36] In fact, an extension of such a framework is what is used in Chapter 3 to address the relation between information disclosure, agency, and know-how.

2.4 Extensions and Connections of Stit Theory

In this section I describe interesting extensions of *BST*, where each extension highlights a connection between stit theory and other disciplines in the literature on applied logic.[37] Since the import-export exchange of ideas with such disciplines was stated as one of the reasons for choosing stit theory as my main tool in the formalization of responsibility (see Chapter 1's Section 1.2), I also discuss these connections and their usefulness. Such a discussion will clarify stit theory's position in the literature of formal philosophy. Once again, an outline of this section should help:

- Subsection 2.4.1 presents an extension of atemporal *BST* with groups—or coalitions—of agents. This extension is the first step toward establishing a connection between *BST* and logics for multi-agent systems.
- Subsection 2.4.2 presents an extension of *BST* with action types, connecting stit theory with dynamic logic, coalition logic, and alternating-time temporal logic.
- Subsection 2.4.3 presents an extension of atemporal *BST* with utilities or payoffs, connecting stit theory with game theory and deontic logic.
- Subsection 2.4.4 addresses extensions of atemporal *BST* with epistemic notions, connecting stit theory with epistemic game theory and epistemic logic.

2.4.1 Extension with Group Agency

In the logic-based analyses of interaction and multi-agent settings, how attitudes of individuals relate to those of groups is doubtlessly one of the focus points. In the case of stit theory, the discussion boils down to the concept of *group agency*. Such a concept was first incorporated into *BST* in Horty's (2001) seminal book, starting a line of research within which considerable progress has been made in recent years (see, for instance, Broersen, 2011a; Broersen et al., 2006b; Duijf, 2018; Herzig & Schwarzentruber, 2008; Lorini, 2013; Lorini et al., 2014; Payette, 2014; Schwarzentruber, 2012; Tamminga, 2013).

BST's semantics for individual agency relies on choices of action that are available at some moment. Therefore, it is natural to base group agency on available *joint choices*. All the works mentioned above consider these joint choices

[37] Exploring these extensions is not gratuitous. Each extension implies enriching *BST* with notions that are closely related to the components of responsibility that recurrently appear in this thesis (see the list of components of responsibility on p. 3).

as distributed processes in a specific relation with the individual choices of the group's members (Alur, Henzinger, & Kupferman, 2002; Halpern & Fagin, 1989). The general agreement among stit theorists is that this specific relation must meet the following requirements: (a) for every group—or coalition—$C \subseteq Ags$ and moment m, the set $\mathbf{Choice}^m_C$ of joint choices of action available to C at m should partition H_m; and (b) these joint choices must refine the intersections of the choices available to the members of C at m: for every moment m and $h \in H_m$, $\mathbf{Choice}^m_C(h) \subseteq \bigcap_{\alpha \in C} \mathbf{Choice}^m_\alpha(h)$ (where $\mathbf{Choice}^m_C(h)$ denotes the cell in the partition $\mathbf{Choice}^m_C$ that includes h).

Now, with the notable exceptions of Broersen (2011a), Duijf (2018), and Payette (2014), most of the mentioned authors agree that, as far as point (b) goes, the other inclusion must also hold, so that $\mathbf{Choice}^m_C = \{\bigcap_{\alpha \in C} \mathbf{Choice}^m_\alpha(h); h \in H_m\}$. Following Schwarzentruber (2012), I use the term *super-additivity* (SA) to refer to the following condition for stit-theoretic models: for group $C \subseteq Ags$, moment m, and $h \in H_m$, $\mathbf{Choice}^m_C(h) \subseteq \bigcap_{\alpha \in C} \mathbf{Choice}^m_\alpha(h)$. Following Duijf (2018, Chapter 1), I use the term *intersection property* (IP) to refer to the condition that ensues when the other inclusion also holds: for every moment m and $h \in H_m$, $\mathbf{Choice}^m_C(h) = \bigcap_{\alpha \in C} \mathbf{Choice}^m_\alpha(h)$. In either case, the elements of $\mathbf{Choice}^m_C$ are interpreted as the joint choices available to C at m, and they provide semantics for a modality of the form $[C]\varphi$, meant to express that coalition C has seen to it that φ. The following definitions make these remarks formal, where I refer to the extension of atemporal *BST* with group notions as *atemporal group stit theory*.[38]

Definition 2.22 (Syntax for atemporal group stit theory). *Given a finite set Ags of agent names and a countable set of propositions P, the grammar for the formal language $\mathcal{L}_G$ is given by*

$$\varphi ::= p \mid \neg\varphi \mid \varphi \wedge \varphi \mid \Box\varphi \mid [C]\varphi,$$

where p ranges over P and C ranges over 2^{Ags}.

$[C]\varphi$ expresses that 'coalition C has seen to it that φ' (and $\langle C \rangle \varphi$ abbreviates $\neg[C]\neg\varphi$). In this setting, $[\alpha]\varphi$ abbreviates $[\{\alpha\}]\varphi$.

Definition 2.23 (Group agency). *For a* bt-*model $\mathcal{M}$ such that $m \in M$ and $h \in H_m$, the semantics for the formulas of $\mathcal{L}_G$ are obtained by extending the recursive definition in Definition 2.3 with the following clause:*

$$\mathcal{M}, \langle m,h \rangle \models [C]\varphi \quad \textit{iff} \quad \textit{for all } h' \in \mathbf{Choice}^m_C(h), \mathcal{M}, \langle m,h' \rangle \models \varphi.$$

[38] Of course, there exist stit logics for group agency whose language also includes the temporal operators G and H (see, for instance, Lorini, 2013). Although they will not play a prominent role in what follows, I label the logics of this form under the term *temporal group stit theory*.

For the sake of completeness, let me address the basic metalogic results for atemporal group stit theory. First shown by Herzig and Schwarzentruber (2008), it is a well-known fact that, when $card(Ags) > 2$, the satisfiability problem of a formula of $\mathcal{L}_{\mathsf{G}}$, with respect to *bt*-models where the intersection property (IP) holds, is undecidable. The proof in Herzig and Schwarzentruber (2008) also shows that atemporal group stit theory over the class of *bt*-models where (IP) holds is not finitely axiomatizable. However, if one considers the bigger class of models given by *bt*-models where only (SA) (super-additivity) holds, then atemporal group stit theory is indeed finitely axiomatizable with respect to such a class, as well as decidable (Broersen, 2008a; Schwarzentruber, 2012).

2.4.2 Extension with Action Types

In the logic-based modelling of action and ability, "stit is not the only game in town" (van Benthem & Pacuit, 2014, p. 292). According to van Benthem and Pacuit, there are two general views on how to model actions available to an agent by means of modal logic. One is the view of *BST*, and the other comes from dynamic logic (Harel, Kozen, & Tiuryn, 2001). The main ingredients of *BST* have already been discussed, so let us turn our attention to the second approach.

Propositional dynamic logic (*PDL*) (M. J. Fischer & Ladner, 1979; Harel, 1984), originated as the propositional fragment of Pratt's (1982) *dynamic logic*. It is based on the idea that actions determine transitions between states of a system. These transitions are typically represented by relations on a set W of possible worlds, where each $w \in W$ represents a state of a system. A transition-relation is labelled by an *action label* or *action type*, so that if all action labels are collected in a set *Types*, then, for each label $a \in Types$, the relation $R_a \subseteq W \times W$ indicates the possible executions of basic action—or basic program—a. Basic programs are then used to construct complex ones according to specific rules.[39]

As for the syntax of *PDL*, for a complex program π, the language $\mathcal{L}_{\mathsf{PDL}}$ of *PDL* includes modalities of the form $[\pi]\varphi$, meant to express that after every possible execution of program π formula φ holds. As for the semantics, the formulas of $\mathcal{L}_{\mathsf{PDL}}$ are evaluated on Kripke structures known as *labelled transition systems* (see, for instance, Alur et al., 2002;Blackburn et al., 2002, Chapter 2). It is with respect to these transition systems that an ontology of action arises. Here, the performance or execution of an action type is seen as an event that changes the world, in the sense that it causes a system to transition from some state to another. For instance,

[39] Thanks to its application to deontic logic (Meyer, 1988) and to philosophy of action (Segerberg, 1992), *PDL* began being considered as a logic of action.

if I raise my arm, I am performing a type of the form 'to raise one's arm,' and the world changes from a state at which my arm is lowered to a state at which my arm is raised.

As far as comparisons go, it is hard to say whether *PDL* is better or worse at formalizing actions than *BST*. *PDL* is an interesting framework with a wide variety of applications (Baltag & Renne, 2016; Harel, 1984; Meyer, 1988; Segerberg, 1992; Troquard & Balbiani, 2019). Furthermore, it has convenient metalogic properties: as shown by Blackburn et al. (2002, Chapters 4 & 6), there is a simple proof system for *PDL* that is sound and complete with respect to labelled transition systems, as well as decidable. Thus, it seems appropriate to say that, in the formalization of actions, favoring *BST* over *PDL* or vice versa depends on the purpose of the formalization. In deciding which logic would be more suitable for any such purpose, it is convenient to bear in mind that the differences between these frameworks can be summarized in two main points (which are related to one another to the extent that one refers to the semantic aspect of the logics and the other to the syntactic one):

- Conceptually, *PDL* treats the execution of an action as a transition from a state to another, where the concept of time is not explicit. In contrast, *BST* treats actions as ways in which agents bring about states of affairs in the world at some moment. In *BST*, agency depends on the effects caused by the actions available to agents.

- Syntactically, *PDL* has a language that includes action types. In contrast, *BST* does not explicitly include terms for actions in its language (see the second bullet point in Remark 2.4). However, unlike *PDL*, *BST* does include terms that refer to agents, as is clear from the exposition in this chapter.

Regardless of these differences, or perhaps because of them, the literature has taken an interest in exploring hybrid systems. The attempts are divided in two main categories: those enriching *PDL* with stit-theoretic notions, and those enriching *BST* with notions from *PDL*. The latter category serves as background for reviewing specific connections between stit theory and logics for multi-agent systems (MAS). Such connections support the claim that one of the biggest advantages of stit theory, at least regarding the formalization of responsibility, is its relation to other theories (see Chapter 1's Section 1.2). Thus, the latter category will be discussed in far more detail than the former, whose basics I briefly describe below.

Incorporating stit-theoretic notions into *PDL*

In *PDL*, the execution of action types is not explicitly linked to agents. Therefore, the main import from stit theory is the notion of agency, seen as a measure of the control that an agent can have over the execution of an action type (van Benthem & Pacuit, 2014). The most common technique for incorporating agency into *PDL* involves using complex action types of the form (α, π), where $[(\alpha, \pi)]\varphi$ is meant to express that φ will be the case after every execution of program π by agent α (Canavotto, 2020; Herzig & Longin, 2004; Lorini & Herzig, 2008; Meyer, van der Hoek, & van Linder, 1999; van Benthem & Pacuit, 2014). A relation $R_{(\alpha,\pi)}$ is what underlies modality $[(\alpha, \pi)]\varphi$ in labelled transition systems, so that $\mathcal{M}, w \models [(\alpha, \pi)]\varphi$ iff for each v such that $wR_{(\alpha,\pi)}v, \mathcal{M}, v \models \varphi$. As mentioned by Canavotto (2020, Chapter 2), unless certain restrictions are applied to $R_{(\alpha,\pi)}$, this notion of agency has the following differences with stit theory's: (a) the available action types do not necessarily partition the set of transitions starting at some state, (b) it is not necessary for all agents to execute an action type for a transition to ensue, and (c) the execution of some type by an agent might not be independent of the execution of another type by another agent.

Incorporating *PDL* notions into *BST*

Just as relations underlie transitions between states in *PDL*, *BST*'s semantics for agency also involves relations, namely the equivalence relations underlying choice-partitions $\mathbf{Choice}^m_\alpha$ (with $\alpha \in Ags$). Therefore, there are natural ways of extending *BST* with action labels in the tradition of *PDL*. For instance, one can incorporate a set of action labels into the models and use it to tag the equivalence relations in such a way that different relations can be tagged with the same label. Indeed, this method has been explored by different authors (see, for instance, Herzig & Troquard, 2006; Horty, 2019; Horty & Pacuit, 2017; Lorini et al., 2014). These extensions, however, are only semantic in nature; the basic syntax of the logic remains unchanged. Below, I present the formal definition of the models that result from incorporating action labels into *bt*-models, following the exposition of Horty and Pacuit (2017):

Definition 2.24 (Labelled *bt*- frames & models)**.** *A tuple of the form* $\langle M, \sqsubset, Ags, \mathbf{Choice}, Tps, Lbl, Exe\rangle$ *is called a* labelled *bt*-frame *iff*

- $\langle M, \sqsubset, Ags, \mathbf{Choice}\rangle$ *is a* bt*-frame (Definition 2.2).*
- *Tps is a set of* action types. *For* $\alpha \in Ags$ *and* $m \in M$, Tps^m_α *denotes the set of action types that are available to* α *at* m.

- *Lbl is a label function that maps action tokens to action types: for $\alpha \in Ags$, $m \in M$, and $L \in$* $\mathbf{Choice}_{\alpha}^{m}$*, $Lbl(L) \in Tps$. For $\alpha \in Ags$, Lbl_{α} will denote a function that maps an index to the action label of the action token performed by α at that index. In other words, for index $\langle m,h\rangle$,* $Lbl_{\alpha}(\langle m,h\rangle) = Lbl(\mathbf{Choice}_{\alpha}^{m}(h))$.

- *Exe is a partial execution function that maps each action type $\tau \in Tps$, $m \in M$, and $\alpha \in Ags$ to a particular action token $Exe_{\alpha}^{m}(\tau) \in$* $\mathbf{Choice}_{\alpha}^{m}$*. Lbl and Exe satisfy the following conditions:*
  - (EL) *For each $\alpha \in Ags$ and index $\langle m,h\rangle$,* $Exe_{\alpha}^{m}(Lbl_{\alpha}(\langle m,h\rangle)) = \mathbf{Choice}_{\alpha}^{m}(h)$.
  - (LE) *For each $\alpha \in Ags$, $m \in M$, and $\tau \in Tps$, if $Exe_{\alpha}^{m}(\tau)$ is defined, then* $Lbl_{\alpha}(Exe_{\alpha}^{m}(\tau)) = \tau$.

A labelled *bt*-model $\mathcal{M}$*, then, is a tuple that results from adding a valuation function $\mathcal{V}$ to a labelled* bt*-frame, where $\mathcal{V} : P \rightarrow 2^{I(M\times H)}$ assigns to each atomic proposition a set of indices.*

Even if the extension with action labels is only semantic, it leads to interesting discussions. A famous one—at least in stit theory—concerns the distinction between action *types* and action *tokens*. Paraphrasing Broersen and Abarca (2018b), the difference between types and tokens is as follows: a token is the single performance of an action by a specific agent at a specific moment; action types, in contrast, refer to categories or patterns of actions, that can be repeated at different moments and that are instantiated in tokens. For instance, when one says 'I opened the window of Fyodor's bedroom at 4 a.m. on Monday,' this is generally seen as an action token. The expression 'to open a window,' in turn, can be thought of as a type. The literature on logics of action knows no consensus as to how specific tokens must be, or how general types must be, so the exact border between the two concepts is ambiguous. What is usually taken for granted, then, is that the more specific the conditions of performance are, the more the action is seen as a token; and the more applicable the term for an action is to many particular instances, the more the action is seen as a type.

A usual assumption among stit theorists is that the cells of choice-partitions in *bt*-models are action tokens rather than types. The reason is that these cells are specific to each point in time and to each agent. Thus, extending the models with types typically accompanies the intent of binding several tokens together under a unifying term. In several recent works (for instance, Horty, 2019; Horty & Pacuit, 2017; Lorini et al., 2014), the purpose of binding tokens together is the same: to model the interplay between knowledge and agency so that the same types of actions are available at indistinguishable moments.[40]

Therefore, labelled *bt*-models are interesting by themselves, and although they narrow the gap between stit theory and *PDL*, there exist two logic-based formalisms that resemble the logic of labelled *bt*-models much more than any actual dynamic logic.[41] These are *coalition logic* (*CL*) (Pauly, 2002) and *alternating-time temporal logic* (*ATL*) (Alur, Henzinger, & Kupferman, 1997; Alur et al., 2002; Goranko & van Drimmelen, 2006).[42]

On the one hand, *CL* and *ATL* share with *PDL* the semantic, conceptual standpoint on action. Both logics treat the performance of actions as events in a transition system. However, neither *CL* nor *ATL* uses action labels in the object language. On the other hand, as exposed by Broersen et al. (2006a, 2006b), there exist clear metalogic relations between appropriate extensions of *BST* and the logics *CL* and *ATL*. To clarify, extending *BST* with action labels (in the models), next-moment operators, group notions, and strategies yields a framework that subsumes both *CL* and *ATL*. Therefore, it is fair to say that *CL* and *ATL* lie somewhere in the middle between *BST* and *PDL*.

To conclude this subsection, it must be mentioned that there are *PDL*-inspired extensions of branching-time logic that are not merely semantic extensions. A technique for incorporating action terms into the syntax of a stit-like framework comes from using propositional constants of the form $do_{\alpha}(\tau)$, standing

[40] This quality is usually referred to as *uniformity of available action types* (UAAT). To the best of my knowledge, the usefulness of incorporating action types into *BST* to account for (UAAT) was first highlighted by Herzig and Troquard (2006)—in the context of formalizing the *epistemic* sense of ability, which is closely related to the concept of *know-how*. However, Herzig and Troquard (2006) managed to circumvent the introduction of action types to *BST* and still succeeded in characterizing (UAAT). Their method was further developed by Duijf et al. (2021), who showed that types are not necessary to formalize either (UAAT) or the epistemic sense of ability. All these topics are prominent throughout this thesis. They are introduced in Chapter 3 and discussed at length in Chapter 4.

[41] By 'actual dynamic logic' I refer to any logic that includes action labels in its models and an explicit treatment of action terms in the object language.

[42] First presented by Pauly (2001), *coalition logic* (*CL*) is a modal logic to formalize *group ability*. Broadly speaking, it provides a logic-based criterion for deciding when a coalition is collectively able to bring about a state of affairs as a particular outcome of a strategic game. Just as in *BST*, *CL*'s notion of (causal) ability is based on the availability of actions (see Subsection 2.2.3). The exact relation between *CL* and *BST* was first formally explored by Broersen et al. (2006b), whose results imply that *CL* can be embedded in an appropriate extension of *BST* (whose language includes the next-moment operator X from xstit theory and group-agency operators $[C]$). For the technical details of the embedding, the reader is referred to Broersen et al. (2006b); Canavotto (2020). Developed by Alur et al. (1997, 2002), *alternating-time temporal logic* (*ATL*) was presented as an extension of *computation tree logic* (*CTL*), a branching-time temporal logic designed to reason about properties of computations in a system. While *CTL* includes modal operators for the universal, resp. existential, quantification over sets of *paths* or *computations*—sequences of states such that each element of the sequence transitions

for 'agent α performs an action of type τ' (see, for instance, Broersen, 2014a; Canavotto, 2020; Herzig & Lorini, 2010). For a language of branching-time logic including the operator $\Box$ for historical necessity and the next-moment operator X, the extension of such a language with these propositional constants yields that one can define modalities for agency so that they work just as stit-theoretic modalities: for a set Tps_α of action types available to agent α, one can define $[\alpha]^X\varphi := \bigvee_{\tau \in Tps_\alpha} (do_\alpha(\tau) \wedge \Box(do_\alpha(\tau) \to X\varphi))$. Thus, frameworks of this kind include actions at the level of the object language, making it is possible to syntactically characterize some intended property of types themselves. This leads to the development of very interesting logics, to say the least (see, for instance, Canavotto, 2020, Chapter 3).

2.4.3 Extension with Utilities & Obligations

As mentioned before, the layouts of choice-partitions in *bt*-frames look very similar to normal-form games. However, there are subtle differences to bear in mind. Normal-form games include payoff functions that assign an individual payoff, or *utility*, to each outcome and agent—where these outcomes are given by full 'strategy' profiles.[43] With this scheme of actions and payoffs, game theory allows us to reason about two questions: (1) how agents *can* act, given their available actions and the actions of other agents; and (2) how agents *would need to* act if their choices were guided by particular combinations of their preferences with varying assumptions for their rationality.[44]

into the next, *ATL* was introduced to reason about *strategies* over such transition systems. The main question was to formalize when a coalition is able to choose and perform a strategy—seen as a set of alternative paths—such that φ is guaranteed to occur at specific states of all the paths belonging to the strategy. Goranko (2001) showed that *CL* is a fragment of *ATL*. This result made it possible to think of a formal relation between *ATL* and stit theory. Broersen et al. (2006a) explored such a relation and gave the essential ideas to embed *ATL* into stit theory, but the technical details for such an embedding are missing in the literature. The interested reader is referred to a preprint of mine (`https://arxiv.org/pdf/2302.07332.pdf`), that shows that *ATL* can indeed be embedded in an extension of *BST* with group strategies.

[43] In game theory, the word 'strategy' is used in different ways in different contexts. In normal-form games, a strategy is a single choice of action available to some player. In extensive-form games, a player's strategy is a function that maps chronologically ordered sequences of that player's actions, each from the game-tree nodes at which it is the agent's turn to act, to a next action of that agent (see Osborne & Rubinstein, 1994, Chapter 6). From here on, I will use the term 'strategy' to refer to the notion as is done in the analysis of extensive-form games (or in *ATL*, for that matter), and I will use the term 'action' to refer to the notion as is done in the analysis of normal-form games. Therefore, in normal-form games each outcome is associated with a full *action* profile.

[44] These questions imply that game theory is closely related to two other philosophical disciplines: *decision theory* and *rational choice theory*, which respectively study the reasoning processes underlying specific choices of agents and how these reasoning processes aggregate to social behaviors in the context of interdependent decision contexts (see Steele & Stefánsson, 2016). These questions also imply

BST shares with game theory the use of models to address question (1) above. A natural extension of *BST*, then, would result from adding payoffs or utilities to *bt*-models in order to accommodate some philosophical purpose. Starting with the proposals of Horty (2001) and of Kooi and Tamminga (2008) (see also Tamminga, 2013; van de Putte, Tamminga, & Duijf, 2017), this kind of extensions has already appeared in the literature, and the philosophical purpose has mostly been the same: the formalization of *optimality* and *obligation*. Since they serve as an important background to much of the work done in Chapters 4 and 6, I address the basic aspects of such extensions below.

Following Duijf's (2018, Chapter 1) exposition, consider the following class of *bt*-frames, resulting from extending the tuples in Definition 2.2 with payoff functions.

Definition 2.25 (Rich Kripke-*stit*-frames). *A tuple* $\left\langle W, Ags, R_\Box, \texttt{Choice}, \{u_\alpha\}_{\alpha \in Ags} \right\rangle$ *is called a* rich Kripke-*stit*-frame *iff*

- $\langle W, Ags, R_\Box, \texttt{Choice} \rangle$ *is a Kripke-*stit*-frame.*
- *For all* $w, v \in W$, *it holds that* $wR_\Box v$.
- *For all* $\alpha \in Ags$, $u_\alpha : W \to \mathbb{R}$ *is a utility function assigning a real value* $u_\alpha(w)$ *to each dynamic world in* W.

As reviewed by Duijf (2018, Chapter 1), rich *bt*-models are very similar to *consequentialist models* (Kooi & Tamminga, 2008) and to *consequentialist choice Kripke models* (Ciuni & Horty, 2014). Ciuni and Horty's and Duijf's correspondence results imply that the class of strategic normal-form games corresponds to a specific sub-class of rich Kripke-*stit*-frames, namely *deterministic* rich Kripke-*stit*-frames.[45]

Thus, there exists a specific connection between *BST* and game theory. This connection brings much conceptual insight on the stit-theoretic modelling of agency, as discussed both by Ciuni and Horty (2014) and by van Benthem and Pacuit (2014). To clarify, one can take to stit theory any intuition—about the decision-making process in interdependent decision contexts—that strategic normal-form games are able to express.[46] In fact, most of this chapter's discussion has already taken advantage, at least implicitly, of this incorporation of ideas. As

that, just as with decision theory, game theory can be thought as having two branches: *descriptive game theory*, used to reason about how agents can act, and *normative game theory*, used to reason about how agents should act if they abode by certain principles.

[45] A rich Kripke-*stit*-frame is called determinisic iff for all $w \in W$, $\bigcap_{\alpha \in Ags} \texttt{Choice}^{\overline{w}}_{\alpha}(w) = \{w\}$.

[46] A good example of this can be found in Chapter 3, where I incorporate into stit theory epistemic notions that come from EGT (*ex ante*, *ex interim*, and *ex post* knowledge).

stated by Ciuni and Horty, stit theory was originally conceived as a logic only about the contribution of agents to changes in the world, but the similarities of its models with game-theoretic ones drove stit theorists to also speak about agents' *choices* or *decisions*, notions that have been central to this chapter's presentation.[47]

It is following this conceptual connection that adding utilities to stit-theoretic models has been used to reason about other philosophical concepts. Before, I mentioned that both Horty's (2001) and Kooi and Tamminga (2008)'s proposals—according to which utility functions can be used to formalize obligations in stit theory—constitute an important background for Chapters 4 and 6. Furthermore, they imply drawing a bridge between stit theory and yet another discipline: *deontic logic.* The intuition behind this bridge lies in normative game theory (see Footnote 44), where assumptions on utilities and on the behavior of decision makers underlie standards for how agents should act in relation to the circumstances of a game.[48] By letting these circumstances describe ethical, moral, or legal situations, for instance, and by adopting a *utilitarian* perspective, the mentioned works adapt game-theoretic tools into stit theory to formalize the deontic concept of *ought-to-do* (Horty, 2001, Chapter 4). The basic idea is that an agent's choices can be compared with each other according to the utilities or payoffs that they lead to, so that the best—or optimal—choices support what the agent ought to have brought about. I refer to the resulting formalism as *act-utilitarian stit theory*, and its detailed exposition can be found in Chapter 4's Subsection 4.2.1.

2.4.4 Extension with Epistemic Notions

In Footnote 9 I mentioned that, according to van Benthem and Pacuit (2014), one can recognize four main stages in choice scenarios: *deliberation*, *decision*, *action*, and *observation*. Although this chapter has touched upon all these stages, for-

[47] According to Ciuni and Horty (2014), game-theoretic ideas were very important for stit theory since the latter's beginning. This is clearly illustrated by Belnap et al. (2001, Chapter 4, pp. 283, 343–344), where the matrix representation of games is used to explain frame condition *independence of agency*, and where a comparison between extensive-form games and *bt*-models is briefly addressed. As for this last point, the definition of an extensive-form game bears plenty similarities with Definition 2.19's *bdt*-frames (for the precise definition of an extensive-form game, the reader is referred to Osborne and Rubinstein (1994, Chapter 6, p. 89)). In fact, the correspondence between concurrent game structures and labelled *bdt*-models in my preprint `https://arxiv.org/pdf/2302.07332.pdf` can be used to establish a correspondence result between extensive-form games and *rich deterministic labelled* bdt-*models*, where these last structures result from adding utility functions u_α (for each $\alpha \in Ags$) to deterministic labelled *bdt*-models so that, for each $\alpha \in Ags$, u_α assigns to each history a utility.

[48] To be more precise, these standards refer to game theory's *solution concepts*, according to which the actions available to an agent in a game can be compared with each other. Thus, 'preferred,' 'better,' 'optimal,' or 'more appropriate' actions arise. Among the most common solution concepts one finds *Nash equilibrium*, *iterated elimination of either strictly or weakly dominated actions*, *iterated elimination of never best responses*, and *regret minimization* (see, for instance, Osborne & Rubinstein, 1994).

malizations have essentially been limited to *action*. What about the other three? Does stit theory have something to say about any of them? Well, stit theory's connection with game theory provides a good context for beginning to answer this question.

Epistemic game theory (EGT) (see Pacuit & Roy, 2017, for an introduction) is a sub-field of game theory that sees decision making in games as similar to what decision theorists know as *choice under uncertainty*. Decision theory states that to choose rationally is to select the best action in light of one's beliefs or one's information about the world. It is a general assumption that agents do not know everything about the world, so that their choices are made under varying degrees of uncertainty. Choosing an action, then, is assumed to involve what agents *know*, what they *believe*, and what they *prefer* about the world. Now, EGT "can be seen as an attempt to bring back the theory of decision making in games to its decision-theoretic roots" (Pacuit & Roy, 2017). Thus, performing an action in a game now involves what agents *know*, what they *believe*, and what they *prefer* about the game. Consequently, EGT brings to the table a formal account of knowledge and beliefs and incorporates it into classic game theory.

Such an incorporation implies studying the aforementioned stages of choice scenarios. Since at the stage of deliberation "we analyze our options and find optimal choices," at the stage of decision "we make up our mind and choose an action of our own," and at the stage of observation the actions of others "get observed" (van Benthem & Pacuit, 2014, p. 309), then these three stages directly concern agents' knowledge, beliefs, and preferences across the decision-making process. Thus, the game-theoretic flavor of stit theory is what set the tone for incorporating epistemic notions into stit theory itself, and this led to a good amount of research over the last two decades. As the literature's ideas in these respects are important for the rest of the thesis, I present their basics below.

Herzig and Troquard (2006) were the first to add epistemic modalities to the basic stit language $\mathcal{L}$, giving rise to what from here on I will refer to as *epistemic stit theory* (*EST*).[49] The relationship between knowledge and agency within stit theory was further explored in a series of papers by Broersen (2008a, 2008b, 2011a). In turn, the last work in this series opened a line of research connecting *EST* with deontic logic (Horty, 2019; Horty & Pacuit, 2017; Lorini et al., 2014), in the same spirit as the connection discussed in the previous subsection.[50]

[49] The goal of Herzig and Troquard (2006) was to use epistemic *bdt*-frames to characterize an interpretation of know-how as the composition of the modalities for historical possibility, for knowledge, and for agency: to know how to see to it that φ was equated to having the possibility of seeing to it that φ at all epistemically indistinguishable indices. A discussion of this proposal is discussed extensively in Chapter 3's Subsection 3.3.4.

[50] The works in this line of research are the main background of Chapters 3, 4, and 5.

As for the syntax of *EST*, it is given by the following definition:

Definition 2.26 (Syntax of *EST*). *Given a finite set Ags of agent names and a countable set of propositions P, the grammar for the formal language $\mathcal{L}_K$ is given by*

$$\varphi ::= p \mid \neg\varphi \mid \varphi \wedge \varphi \mid \Box\varphi \mid [\alpha]\varphi \mid K_\alpha\varphi,$$

where p ranges over P and α ranges over Ags.

In this language, $\Box\varphi$ and $[\alpha]\varphi$ have the same meanings as in *BST*; $K_\alpha\varphi$, in turn, expresses that 'agent α knows φ.' As for the semantics, the structures on which the formulas of $\mathcal{L}_K$ are evaluated are based on what I refer to as *epistemic branching-time frames*. The intuition behind this kind of frames is that each moment is a scenario of choice under uncertainty. In other words, agents make choices at points in time while having *incomplete* and *imperfect information* about the actual evolution of the world.[51]

Definition 2.27 (*Ebt*-frames & models). *A tuple* $\left\langle M, \sqsubset, Ags, \mathbf{Choice}, \{\sim_\alpha\}_{\alpha \in Ags}\right\rangle$ *is called an* epistemic branching-time frame *(*ebt-*frame for short) iff*

- $\langle M, \sqsubset, Ags, \mathbf{Choice}\rangle$ *is a* bt-*frame (Definition 2.2).*
- *For all $\alpha \in Ags$, $\sim_\alpha$ is an equivalence relation on the set of indices, meant to represent the epistemic indistinguishability relation for α.*

An ebt-*model $\mathcal{M}$, then, is a tuple that results from adding a valuation function $\mathcal{V}$ to an* ebt-*frame, where $\mathcal{V} : P \to 2^{I(M \times H)}$ assigns to each atomic proposition a set of indices.*

Definition 2.28 (Evluation rules for *EST*). *Let $\mathcal{M}$ be an* ebt-*model. The semantics on $\mathcal{M}$ for the formulas of $\mathcal{L}_K$ are obtained by extending the recursive definition in Definition 2.3 with the following clause:*

$$\mathcal{M}, \langle m, h\rangle \models K_\alpha\varphi \quad \textit{iff} \quad \textit{for all } \langle m', h'\rangle \textit{ s. t. } \langle m, h\rangle \sim_\alpha \langle m', h'\rangle, \mathcal{M}, \langle m', h'\rangle \models \varphi.$$

As for the logic-based properties of this kind of knowledge, basing the semantics of $K_\alpha\varphi$ on an equivalence relation implies that K_α is a **S5** modal operator. Thus, the validity of schema (K) implies that agents are idealized logical thinkers with so-called *logical omniscience*: they know all the logical consequences of their

[51] Ågotnes, Goranko, Jamroga, and Wooldridge (2015, emphasis in original) wrote: "[i]n game theory, two different terms are traditionally used to indicate lack of information: 'incomplete' and 'imperfect' information. Usually, the former refers to uncertainties about the game structure and rules, while the latter refers to uncertainties about the history, current state, etc. of the specific *play of the game.*"

knowledge.[52] The validity of (T) means that knowledge is *factive*: only truths can be known. The validity of (4) means that agents have *positive introspection*: if they know something, then they know that they know it. Finally, the validity of (5) implies that agents have *negative introspection*: if they do not know something, then they know that they do not know it.[53]

The topics that are common to all the works that first added epistemic modalities to *BST* have to a certain extent defined *EST*. These topics are the following:[54]

- *Knowingly doing*: interpreting moments of *bt*-models as choice scenarios with incomplete and imperfect information led to a distinction between what agents do unknowingly and what agents knowingly do. Concepts first introduced by Broersen (2008a), the literature has somewhat settled on the stit-theoretic interpretation of knowingly (and unknowingly) doing. On the one hand, at index $\langle m,h\rangle$ agent α has knowingly seen to it that φ iff $\mathcal{M},\langle m,h\rangle \models K_\alpha[\alpha]\varphi$—that is, iff at $\langle m,h\rangle$ α knew that it has seen to it that φ. On the other hand, at $\langle m,h\rangle$ α has unknowingly seen to it that φ iff $\mathcal{M},\langle m,h\rangle \models [\alpha]\varphi \wedge \neg K_\alpha[\alpha]\varphi$—that is, iff at $\langle m,h\rangle$ has seen to it that φ but did not know this.[55]

- *Epistemic sense of ability* and *know-how*: just as in the above item, adding uncertainty to *bt*-models allows us to distinguish an epistemic kind of ability from mere causal ability (see Subsection 2.2.3). Thus, the ability to knowingly see to it that some state of affairs ensues—as opposed to the ability of just seeing to it that it occurs—has been referred to as the epistemic sense of ability (Horty & Pacuit, 2017). In terms of formulas of $\mathcal{L}_{\mathsf{K}}$, at $\langle m,h\rangle$ α was able in the epistemic sense to see to it that φ iff $\mathcal{M},\langle m,h\rangle \models \Diamond K_\alpha[\alpha]\varphi$—that is, iff at $\langle m,h\rangle$ it was historically possible for α to knowingly see to it that φ. Stit theorists have often linked this epistemic sense of ability with the

[52] This issue is controversial, to say the least, and the reader is referred to Stalnaker (1991) for an interesting exposition of the problem.

[53] Negative introspection with respect to knowledge is a highly controversial property in the philosophical literature. It seems unlikely that someone would know that they do not know something. The reader is referred to Lenzen (1979) for a thorough examination of the matter. That said, one should also keep in mind that, although much more appealing than its negative counterpart, positive introspection has also been taken to test. Williamson (2002), for instance, advocated interesting reasons for rejecting the principle.

[54] All these topics are thoroughly discussed, and further formalized, in Chapter 3.

[55] Broersen (2008a, p. 53, emphasis in original) wrote that the things that an agent does unknowingly "vastly outnumber the things an agent knows he does. For instance, by sending an email, I may enforce many, many things I am not aware of, which are nevertheless the result of me sending the email. All these things I do *unknowingly* by knowingly sending the email."

so-called know-how (also known as practical or procedural knowledge).[56] Although some authors might disagree with a complete identification of know-how with the epistemic sense of ability, it is hard to disagree with the argument that in order to know how to do something one must be able in the epistemic sense to do it.

- *Knowledge across the stages of information disclosure*: the stages of choice scenarios that were mentioned before (deliberation, decision, action, and observation) admit different types of knowledge, each progressively refining the one from the previous stage. EGT has come to distinguish three kinds of knowledge therein (according to the process of information disclosure in multi-agent decision making): *ex ante* knowledge, that concerns the information that is available to agents regardless of their choices; *ex interim* knowledge, concerning the knowledge that is private to an agent after choosing an action but before having information about the concurrent choices of other agents; and *ex post* knowledge, concerning the information that is disclosed after everybody executes their choices. A stit-theoretic formalization of these kinds of knowledge is presented in Chapter 3.

- *Uniformity*: adding indistinguishability relations to *bt*-models led to reasoning about agency at indistinguishable states. The idea, common to EGT and the epistemic extensions of logics for multi-agent systems,[57] that an agent should have the same available actions at indistinguishable states is known as *uniformity*.[58] In the literature on *EST*, there have been two ways of modelling uniformity: (a) through *uniformity of available action types*, for which

[56] According to Duijf (2018, Chapter 3), Fantl (2008) drew the outlines of know-how by distinguishing it from two other kinds of knowledge: *knowledge by acquaintance* and *propositional knowledge* (know-that). Setting aside the concept of knowledge by acquaintance, Duijf proposed that the essential difference between know-how and know-that lies in the content that they take. Procedural knowledge takes actions as content, and propositional knowledge takes propositions as content. I agree with this interpretation, which identifies the know-how studied here with Wang's (2018) goal-directed know-how, for instance. Apropos, Ryle (2009) introduced a debate as to whether know-how can be reduced to know-that or not, where *intellectualists* think that it can and *anti-intellectualists* think that it cannot.

[57] The epistemic extension of *ATL* is a famous member in this category of epistemic logics for multi-agent systems. Known in the literature as *alternating-time temporal epistemic logic* (*ATEL*), this framework was developed in a series of papers to reason about imperfect and incomplete information in concurrent game structures (see Ågotnes, 2006; Ågotnes et al., 2015; Jamroga & Ågotnes, 2007; van der Hoek & Wooldridge, 2002). Thus, its analysis is similar to that of EGT's extensive-form games with information sets.

[58] According to Horty and Pacuit (2017), the use of the term 'uniformity' in this context is due to van Benthem (2001). The concept is familiar in EGT, where the condition of *uniform strategies* is captured by the fact that a strategy for an agent is defined as a function from that agent's information sets to actions, rather than from game states to actions. The term *uniform strategies* is also used in *ATEL*, to refer to the analogous property that agents must be able to perform the same in-the-long-run strategies

the practice of tagging different action tokens under the same type allows us to describe uniformity across indistinguishable indices in labelled *bt*-models (see Definition 2.24); and (b) through *uniformity of historical possibility*, that establishes that the same knowledge should be historically possible at indistinguishable indices. Both conditions are further explored in Chapters 3 and 4, and they were shown to correspond to one another by Duijf et al. (2021).[59]

Of course, all these topics imply drawing a bridge between stit theory and a huge field in formal philosophy known as *formal epistemology*. More precisely, they imply the incorporation into stit theory of ideas from the—also huge—field known as *epistemic logic*.[60]

Now, although the stit-theory community has shown a considerable interest in exploring the influence of knowledge on agency over the last two decades, there are relatively few extensions of *BST* with *belief* modalities. To the best of my knowledge, such an extension has been previously addressed only by Wansing (2006a), by Broersen (2011c), and in my recent joint work with Jan Broersen (Abarca & Broersen, 2021a). The first study incorporated beliefs into *BST* to analyze the concept of *doxastic voluntarism*—a philosophical premise by which agents actively decide to acquire certain beliefs. The second one added probabilities to *bt*-models to represent the degrees of belief that an agent can have regarding the bringing about of some state of affairs as an effect of one of its available actions. Finally, Abarca and Broersen (2021a) picked up the connection between game theory (EGT, in particular) and act-utilitarian stit theory to explore the interplay between beliefs, agency, and obligations. In fact, the conclusions of Chapters 3, 4, and 6 explore a version of this last proposal, to account for the importance of belief as a component of responsibility.

2.5 Conclusion

This long chapter was devoted to a logic-based characterization of agency. The basic syntactic and semantic aspects of stit theory—the logic of action lying at

at epistemically indistinguishable states (where these strategies should assign the same choices to indistinguishable states). A well-known issue is that uniform strategies lead to complications in *ATEL* (see Herzig & Troquard, 2006, for a brief discussion on the matter).

[59] To clarify, in our joint paper (Duijf et al., 2021) we show that these two conditions can be thought of as corresponding conditions in two different classes of stit models: epistemic labelled *bt*-models, on the one hand, and *ebt*-models, on the other.

[60] For a good introduction to epistemic modal logic, the reader is referred to standard textbooks on the matter (Blackburn et al., 2002; Fagin et al., 1995; Halpern & Fagin, 1989; Hintikka, 1962).

the heart of this thesis—were widely discussed and illustrated, with no little emphasis on the benefits of using *BST* in a potential formalization of responsibility. Similarly, the logic-based and metalogic properties of *BST*, as well as its connections with several frameworks in the literature on applied logic, were thoroughly reviewed. Such discussions, illustrations, logic-based/metalogic properties, and connections were presented to provide a sturdy conceptual background for the succeeding chapters.

"The performance is sometimes masterful, extremely clever, but the control of the actions, their source, is deranged and depends on various morbid impressions," says the suspicious Zossimov, in Dostoevsky's *Crime and Punishment*. By now, we have covered a lot of ground on agency, discussing a conceptualization in terms of choice and performance of actions. But what about the *sources* of agency? And, perhaps more importantly, what about the implications that these sources of agency have on agency itself? In my study of responsibility, the other components in the decomposition's list (p. 3)—knowledge, beliefs, intentions, and obligations)—make up particular instances of such sources; they are some of those 'impressions' that Zossimov talks about, on which the choice and performance of actions depends. The rest of this thesis, then, is devoted to the sit-theoretic analysis of what these sources tell us about agency and about responsibility.

3

Agency & Knowledge

'... to shed blood in all conscience *is to my mind more horrible than if bloodshed were officially, legally permitted'*

Fyodor Dostoevsky, *Crime and Punishment*

'But I didn't realise what I'd done till I heard the sound. Like somebody drowning. Screaming under water. I handed the knife to Dick. I said, "Finish him. You'll feel better." '

Truman Capote, *In Cold Blood*

... they concluded at once that the crime itself could not have occurred otherwise than in some sort of temporary insanity, including, so to speak, a morbid monomania of murder and robbery, with no further aim or calculation of profit.

Fyodor Dostoevsky, *Crime and Punishment*

3.1 Introduction

In April, 1965 Richard Hickock and Perry Smith were executed by hanging at the Lansing Correctional Facility, in Kansas. Their sentence came as a result of a trial that found them guilty of murdering—*in cold blood*—four members of a well-liked family—the Clutters—6 years before. Although it was proven that Hickock and Smith were responsible for the murders, at some point during a long scheme of appeals a case was made that the defendants were afflicted by a psychological impediment that prevented them from *knowing* what they were doing when they shot the Clutters. Under the *M'Naghten Rule*—that prevailed in Kansas law at the time—the only way of pleading 'not guilty by reason of insanity' was with previous medical certification of an acute mental disease forbidding the defendant to "know the nature of their act." In the case of Hickock and Smith, such a mental disease was dismissed. The appeals failed, and the defendants were sentenced to death.

This grim account implies that knowledge plays an important role in the legal, real-life ascription of responsibility to criminals. In turn, for logicians studying agency and obligation, there is little debate about the intuition that responsibility has an important epistemic component. This chapter, then, is devoted to the introduction of epistemic notions to basic stit theory (*BST*), an essential step in building my formal theory of responsibility.

Recent years have seen much interest in the role of knowledge when formalizing responsibility, precisely to deal with situations like Hickock and Perry's example. Aiming to model the degrees of culpability in juridical systems, Broersen (2008a, 2011a) introduced a stit logic of *knowingly doing*. Afterwards, a comprehensive, stit-theoretic study of three kinds of game knowledge (*ex ante, ex interim,* and *ex post*) was provided by Lorini et al. (2014), whose goal was to formalize both responsibility attribution and attribution-emotions related to responsibility (like guilt, or blame). Horty and Pacuit (2017), for their part, added epistemic modalities to atemporal *BST* to formalize both *ex interim* knowledge and an epistemic sense of ability (or know-how). Even more recently, Horty (2019) used such epistemic modalities to create a logic of epistemic obligations. Rather than including all this in a section for related work, I mention it now because this chapter has two main objectives, and both depend on the existing literature on *epistemic stit theory* (*EST*) (see Chapter 2's Subsection 2.4.4):

(i) I want to point out the areas for which the mentioned works overlap, clarifying their differences and their potential shortcomings.

(ii) I want to offer new stit-theoretic formalizations for the notions of *ex ante, ex interim*, and *ex post* knowledge, as well as of *know-how*, targeting components which I believe are central to any analysis of responsibility.

Within epistemic game theory (EGT) and epistemic logic, there is some degree of agreement regarding the four kinds of knowledge mentioned above. In broad terms, their characteristics are the following:

- ***Ex ante knowledge*** concerns information that at a given moment is available to agents regardless of their choices of action. It is commonly thought of as knowledge that agents have before they perform any such choice.
- ***Ex interim* knowledge** concerns information that is private to an agent after choosing an action but before having information about the concurrent choices of other agents. In other words, it includes facts for which an agent knows that they will occur after the choice and performance of one of its actions, independently of what other agents choose.
- ***Ex post* knowledge** concerns information that becomes available after all agents execute their choice.
- **Know-how** concerns an agent's *epistemic sense of ability* (Horty & Pacuit, 2017). By 'epistemic sense of ability' I mean an agent's ability to knowingly perform an action that will bring about a specific outcome. Thus, the version of know-how that I work with is closely related to the possibility of knowingly doing something.

The first three categories are standard in EGT. They refer to an agent's uncertainty during the stages of information disclosure in the decision-making process. Know-how, however, is not explicitly related to these stages. There are two main reasons for including it in the present discussion, then. First, I want to explore the connection between knowledge and ability. Such a connection is unavoidably linked to know-how (see, for instance, Broersen, 2011a; Duijf, 2018; Herzig & Troquard, 2006; Lorini et al., 2014). Secondly, and perhaps more importantly, I focus on knowledge's influence on ability not just for the sake of it; the goal is to aid in the formalization of responsibility. As established by Broersen (2011a) (see also Duijf, 2018; Lorini, 2013), and as illustrated both by Hickock and Smith's example and by the example that I will present in Section 3.2, know-how is very relevant in this matter.[1]

Now, to express the four epistemic notions in clear terms, I use an extension of basic xstit theory (see Chapter 2's Sub-subsection 2.3.3.1). More precisely, the language of this chapter's logic extends the atemporal fragment of *BST*'s language $\mathcal{L}$ (see Definition 2.1, p. 28) with operators for 'next' and 'last' moments and with operators for individual knowledge. As for the semantics, the formulas

[1] The intuition is that someone can be acquitted of knowingly bringing about an undesirable outcome if they did not know how to prevent that outcome. Even if they knew that it was possible to prevent that outcome, maybe they still did not know how to refrain from bringing it about (in the sense that it was impossible for them to either knowingly refrain from doing so or to knowingly bring about a different outcome). Clearly, these are good arguments for excusing them.I return to these points when discussing the version of know-how introduced here (see Subsection 3.3.4).

are evaluated on a special class of *bdt*-frames (see Definition 2.19, p. 57), namely rootless *bdt*-frames including indistinguishability relations for agents. Thus, the resulting logic—which I refer to as *epistemic xstit theory with the grand coalition* (*EXST*)—is an extension of Broersen's (2008a) *Xstit* and an epistemic extension both of Lorini and Sartor's (2016) logic and of a fragment of Payette's (2014) logic. In contrast to Broersen's and Payette's approaches, here (a) group agency is defined in terms of the intersections of individual actions,[2] and (b) both actions with instantaneous effects and actions that take effect at next indices are accounted for. An outline of this chapter is included below.

- Section 3.2 presents an example to illustrate the expressive power of *EXST*. This example helps in the formal introduction of *EXST*'s syntax and semantics.

- Section 3.3 puts forward working definitions for *ex ante*, *ex interim*, and *ex post* knowledge, and for know-how. The section discusses previous interpretations of these notions in the sit-theoretic literature and presents novel *EXST*-based characterizations. These characterizations are then compared with the previous interpretations, using the example of the preceding section for illustration purposes.

- Section 3.4 introduces a Hilbert-style proof system for *EXST* and addresses its soundness & completeness results, sith respect to Kripke models.

- Section 3.5 (the conclusion) explores an extension of *EXST* with another important epistemic notion: belief. Furthermore, an initial characterization of informational responsibility is mentioned.

3.2 An Example Involving Four Kinds of Knowledge

To illustrate the use of *EXST* in the study of multi-agent scenarios of choice under uncertainty—where the four kinds of knowledge play important roles—this section will be presenting a fun (albeit a bit complex) example. For the sake of clarity, let me first introduce the syntax and semantics of *EXST*.

Definition 3.1 (Syntax of *EXST*). *Given a finite set Ags of agent names and a countable set of propositions P, the grammar for the formal language $\mathcal{L}_{KX}$ is given by*

$$\varphi ::= p \mid \neg\varphi \mid \varphi \wedge \varphi \mid \Box\varphi \mid X\varphi \mid Y\varphi \mid [\alpha]\varphi \mid [Ags]\varphi \mid K_\alpha\varphi,$$

[2] That group agency is defined in terms of the intersections of members' actions is a condition that Payette referred to as "complete distributed group action" (Payette, 2014, p. 602). I previously referred to it as the *intersection property* (IP) (see the discussion on p. 60).

where p ranges over P and α ranges over Ags.

In this language, $\Box\varphi$, $[\alpha]\varphi$, and $X\varphi$ have the same meanings as in basic xstit theory (see Definition 2.21, p. 58); $[Ags]\varphi$ expresses 'the grand coalition Ags has seen to it that φ,' and $K_\alpha\varphi$ expresses 'agent α knows φ.' It is important to emphasize that—just as in all the other chapters of this thesis—the description of the stit-theoretic modalities will follow *my interpretation* of the semantics (see the discussion on p. 34 and Remark 2.4). Therefore, when specifying the points of evaluation for the formulas—the indices in *bt*-models—I take it that at those indices states of affairs are definitive. Because of this, I use the present-perfect tense for the description of modality $[\alpha]\varphi$ and say that 'at index $\langle m, h\rangle$ α has seen to it that φ.' To be consistent, I will use the past tense for modalities $\Box\varphi$ and $K_\alpha\varphi$ and say that 'at index $\langle m, h\rangle$ φ was settled φ,' and that 'at index $\langle m, h\rangle$ α knew φ.'

Observe that $\mathcal{L}_{\mathsf{KX}}$ is built with the instantaneous-stit operators $[\alpha]$ and $[Ags]$. Reasons for including these operators are that (a) they allow us to represent agents' available choices in a clear way and (b) the resulting language simplifies the axiomatization for the logic.[3] However, *when talking about the four kinds of agentive knowledge*, the discussion will be restricted to actions that take effect at next indices. Such actions are characterized with formulas of the form $[\alpha]X\varphi$ and $[Ags]X\varphi$. Therefore, both in the present section and in the next one, the fragment of $\mathcal{L}_{\mathsf{KX}}$ that includes only formulas of this type is used. I abbreviate the combination $[\alpha]X$ by $[\alpha]^X$ and the combination $[Ags]X$ by $[Ags]^X$.

As for the semantics, the structures on which the formulas of $\mathcal{L}_{\mathsf{KX}}$ are evaluated are based on what I call *epistemic (rootless) branching-discrete-time frames*.[4]

Definition 3.2 (Epistemic (rootless) branching-discrete-time frames & models)**.** *A tuple* $\left\langle M, \sqsubset, Ags, \mathbf{Choice}, \{\sim_\alpha\}_{\alpha\in Ags}\right\rangle$ *is called an* epistemic rootless branching-discrete-time frame *(*ebdt*-frame for short) iff*

- $\langle M, \sqsubset, Ags, \mathbf{Choice}\rangle$ *is a* bdt*-frame (Definition 2.19, p. 57), where recall that this implies that condition* (TD) time-discreteness *is met: for all $m \in M$ and $h \in H_m$, there exists a unique moment m^{+h} such that $m \sqsubset m^{+h}$ and $m^{+h} \sqsubseteq m'$ for every $m' \in h$ such that $m \sqsubset m'$. For $m \in M$ and $h \in H_m$, m^{+h} is known as the* successor of m along h*. For index $\langle m, h\rangle$, I refer to $\left\langle m^{+h}, h\right\rangle$ as the* successor of $\langle m, h\rangle$ *or as* $\langle m, h\rangle$*'s* next index*.*

[3] I support claim (a) after introducing the semantics for the formulas. I support claim (b) in Section 3.6, Footnote 38.

[4] As implied in Chapter 2's Subsection 2.4.2, these structures are similar to alternating-time temporal logic's concurrent game structures and to extensive-form games.

- *For all $m \in M$ and $h \in H_m$, there exists a unique moment m^{-h} such that $m^{-h} \sqsubset m$ and $m' \sqsubseteq m^{-h}$ for every $m' \in h$ such that $m' \sqsubset m$. For $m \in M$ and $h \in H_m$, m^{-h} is known as the* predecessor of m along h. *For index $\langle m, h\rangle$, I refer to $\left\langle m^{-h}, h\right\rangle$ as the* predecessor of $\langle m, h\rangle$ *or as $\langle m, h\rangle$'s* previous index.[5]

- *For $m \in M$ and $h \in H_m$, $\mathbf{Choice}^m_{Ags}(h)$ denotes the intersection $\bigcap_{\alpha \in Ags} \mathbf{Choice}^m_\alpha(h)$. Thus, $\mathbf{Choice}^m_{Ags} := \left\{\mathbf{Choice}^m_{Ags}(h); h \in H_m\right\}$, which is the partition of actions available to the grand coalition.*

- *For all $\alpha \in Ags$, $\sim_\alpha$ is an equivalence relation on the set of indices, representing the epistemic indistinguishability relation for α. At this point, the only extra condition imposed on these relations is*
  - (NoF) No forget*: for all $\alpha \in Ags$, $m \in M$, and $h \in H_m$, if $\left\langle m^{+h}, h\right\rangle \sim_\alpha \langle m_*, h_*\rangle$, then $\langle m, h\rangle \sim_\alpha \left\langle m_*^{-h_*}, h_*\right\rangle$.*

An ebdt*-model $\mathcal{M}$, then, consists of the tuple that results from adding a valuation function $\mathcal{V}$ to an* ebdt*-frame, where $\mathcal{V} : P \to 2^{I(M \times H)}$ assigns to each atomic proposition a set of indices (recall that P is the set of propositions in $\mathcal{L}_{\mathsf{KX}}$).*

Ebdt-models allow us to provide semantics for the formulas of $\mathcal{L}_{\mathsf{KX}}$:

Definition 3.3 (Evaluation rules for *EXST*)**.** *Let $\mathcal{M}$ be an* ebdt*-model. The semantics on $\mathcal{M}$ for the formulas of $\mathcal{L}_{\mathsf{KX}}$ are defined recursively by the following truth conditions, evaluated at index $\langle m, h\rangle$:*

$$
\begin{array}{lll}
\mathcal{M}, \langle m, h\rangle \models p & \textit{iff} & \langle m, h\rangle \in \mathcal{V}(p) \\
\mathcal{M}, \langle m, h\rangle \models \neg\varphi & \textit{iff} & \mathcal{M}, \langle m, h\rangle \not\models \varphi \\
\mathcal{M}, \langle m, h\rangle \models \varphi \wedge \psi & \textit{iff} & \mathcal{M}, \langle m, h\rangle \models \varphi \textit{ and } \mathcal{M}, \langle m, h\rangle \models \psi \\
\mathcal{M}, \langle m, h\rangle \models \Box\varphi & \textit{iff} & \textit{for all } h' \in H_m, \mathcal{M}, \langle m, h'\rangle \models \varphi \\
\mathcal{M}, \langle m, h\rangle \models X\varphi & \textit{iff} & \mathcal{M}, \left\langle m^{+h}, h\right\rangle \models \varphi \\
\mathcal{M}, \langle m, h\rangle \models Y\varphi & \textit{iff} & \mathcal{M}, \left\langle m^{-h}, h\right\rangle \models \varphi \\
\mathcal{M}, \langle m, h\rangle \models [\alpha]\varphi & \textit{iff} & \textit{for all } h' \in \mathbf{Choice}^m_\alpha(h), \mathcal{M}, \langle m, h'\rangle \models \varphi \\
\mathcal{M}, \langle m, h\rangle \models [Ags]\varphi & \textit{iff} & \textit{for all } h' \in \mathbf{Choice}^m_{Ags}(h), \mathcal{M}, \langle m, h'\rangle \models \varphi \\
\mathcal{M}, \langle m, h\rangle \models K_\alpha\varphi & \textit{iff} & \textit{for all } \langle m', h'\rangle \textit{ s. t. } \langle m, h\rangle \sim_\alpha \langle m', h'\rangle, \\
 & & \mathcal{M}, \langle m', h'\rangle \models \varphi.
\end{array}
$$

Satisfiability, validity, and general validity are defined as usual.

[5] Observe that the definitions of m^{-h} and m^{+h}, coupled with the fact that histories are linearly ordered, implies that, for all $m \in M$ and $h \in H_m$, $\left(m^{-h}\right)^{+h} = m$ and $\left(m^{+h}\right)^{-h} = m$.

Following my interpretation of stit theory, I consider that at an index an agent has already chosen an action and executed it (under some uncertainty). Here, I focus on the effects that such an action has at next indices. That said, it is important to stress that the semantics in Definition 3.3 admit actions with instantaneous effects, with modalities $[\alpha]\varphi$ and $[Ags]\varphi$. As mentioned before, one of the reasons for including such modalities is that their semantics allow us to *explicitly* represent the choices that are available to agents at a given moment (the cells of partitions $\mathbf{Choice}^m_\alpha$ and $\mathbf{Choice}^m_{Ags}$).

The decision to focus only on actions that take effect at next indices comes from the intuition that the chronological dimension of *bt*-frames helps in characterizing the relations between the four kinds of knowledge, especially those concerning *ex post* knowledge. If this claim is a bit obscure, perhaps a more convenient way to think about *ebdt*-models is the following: at index $\langle m, h\rangle$ all agents have performed one of their available actions, namely the one that they chose. Even though these choices may have had instantaneous effects, in this chapter's discussions we can dismiss these effects as those that are inherent to the ongoing choice and performance of an action. In contrast, the *meaningful* effects of actions, meaning those consequences in the world that transcend their mere choice and performance, are those that hold at next indices.

We are ready to present and dissect this section's example. Inspired by the film *Mission: Impossible 6*, this example can be thought of as a variation of Horty and Pacuit's (2017) ones to analyze the epistemic sense of ability.[6]

Example 3.4. *A bomb squad consisting of three members*—Ethan, Luther, *and* Benji—*faces a complex bomb situation. Terrorists have threatened to blow up a facility using two bombs that are remotely connected to each other. Let me call these bombs L and B, respectively (the reason for choosing these labels is that* Luther *will deal with bomb L, and* Benji *will deal with bomb B). As background information, the squad knows the following facts:*

[6] As the reader will see, the example is somewhat involved, and there are three main reasons for having such a complex layout:

(a) I want to illustrate the flexibility of stit-theoretic models in formalizing uncertainty about actions. As such, the example includes instances both of sequential action (previous and succeeding actions at different moments of the same history) and of concurrent action (different agents making their respective choice at the same moment).

(b) I want to study the four kinds of knowledge with the same example.

(c) I want to model a situation that can be interpreted in a context of responsibility attribution. Therefore, it should involve undesirable, resp. desirable, outcomes for which agents can deserve blame, resp. praise, according to (a) their available choices, (b) the knowledge that they had before, while, and after these choices, and (c) their know-how.

- *If the squad defuses one bomb before the other, then the latter is programmed to set off. This means that the squads needs to synchronize their actions.*

- *Each bomb has its own detonator, and each detonator includes a mechanism that makes it possible to cancel the detonation. Let me denote the action of activating both mechanisms by D_{LB}, and the action of* only *activating the mechanism in the detonator for bomb L, resp. B, by D_L, resp. D_B. The terrorists who planted the bombs have triggered both detonators, so there is a countdown to detonation. If the countdown ends, both bombs go off.*

- *Each bomb has a central system that controls it, with two main wires: a red one and a green one. Let me denote the action of cutting the red wire of bomb L, resp. B, by R_L, resp. R_B, and the action of cutting the green wire of bomb L, resp. B, by G_L, resp. G_B.*

- *For the squad to defuse the two bombs at the same time, they need to first deal with the mechanisms in the detonators* and afterwards *synchronize a specific cutting of wires. For this last task, the squad has figured out that there are only three ways to successfully defuse both bombs:*

  (i) *If the squad manages to activate both mechanisms, then afterwards they need to simultaneously cut the red wires of both bombs' central systems. For the sake of clarity—and being informal—we can summarize this by writing that $D_{LB} + (R_L \textit{ and } R_B) = safety$.*

  (ii) *If they manage to activate the mechanism in bomb L's detonator but not in bomb B's, then afterwards they need to simultaneously cut the red wire of L and the green one of B. Informally, $D_L + (R_L \textit{ and } G_B) = safety$.*

  (iii) *The reverse situation of the above item in the case that they only manage to activate the mechanism in bomb B's detonator. Informally, $D_B + (R_B \textit{ and } G_L) = safety$.*

- *If neither mechanism is activated, then both bombs go off. If the combinations mentioned above are not met exactly, then one of the bombs goes off. Summing up, cutting the red wire of a bomb without previous activation of its detonator's mechanism makes it go off; cutting the green wire with previous activation of the mechanism also makes it go off. If any of the bombs goes off, the explosion is so powerful that it is impossible to ascertain which bomb went off or if both did.*

After the countdown starts, Ethan *is commissioned with the task of retrieving the detonators—to activate their mechanisms.* Luther *and* Benji *must afterwards synchronize the cutting of wires in their respective bombs. A malfunction in the squad's telecom gear* *causes for them to lose all communication with each other, so that* Luther *and* Benji *do not know whether* Ethan *has retrieved one detonator, both of them, or none. Regardless, they know that they need to synchronize the cutting of the two bombs' main wires just before the countdown ends.*

To collate cases for which agents' knowledge has consequences in responsibility attribution, I explore the following alternatives for the example's outcome:

(a) Unbeknownst to Luther *and* Benji, Ethan *succeeds in retrieving only the detonator for bomb B.* Luther *and* Benji *synchronize the cutting of wires, and, since it is statistically better for both to cut the red wires, they do so. Bomb L goes off.*

(b) Unbeknownst to Benji, Luther *finds out what* Ethan *did. However,* Luther *is actually an undercover associate of the terrorists, so he decides to go on with the cutting of the red wire, causing bomb L to go off.*

The goal is to formalize Example 3.4 using *EXST*. Thus, one can illustrate instances of *ex ante* knowledge: *what is the information for* Luther *and* Benji *before they have to choose a course of action?*; *ex interim* knowledge: *do* Luther *and* Benji *knowingly choose to set off a bomb?*; ability in the epistemic sense: *in which cases can* Luther *and* Benji *knowingly choose to set off a bomb?*, and *ex post* knowledge: *what is the information after the decisions have been made and executed?*. Therefore, consider the *ebdt*-model $\mathcal{M}$ in Figure 3.1.

Here, $Ags = \{Ethan, Luther, Benji\}$, and m_1–m_{21} are moments, where $\sqsubset$ is defined so as to be represented by the diagram. Histories h_1–h_{16} represent the different possibilities in which the world can evolve from the moment the bomb squad sets out to defuse the bombs onward. For both the alternative endings (case a and case b), the set of moments can be divided in three levels according to the chronological dimension of $\mathcal{M}$:

- *Bottom level:* at the bottom level we find moment m_1. In *ebdt*-models all the agents get to choose from their available actions at every moment. Since *Luther* and *Benji* cannot actually choose anything at m_1, $\mathbf{Choice}^{m_1}_{Luther} = \mathbf{Choice}^{m_1}_{Benji} = \{H_{m_1}\}$. Thus, only the actions available to *Ethan* are relevant: $\mathbf{Choice}^{m_1}_{Ethan} = \{D_{LB}, D_L, D_B, F\}$, where
  - D_{LB} stands for the action of 'activating the mechanism in both detonators.'
  - D_L (D_B) stands for the action of '*only* activating the mechanism in the detonator for bomb L (B).'
  - F stands for the action of 'failing to secure a detonator.'

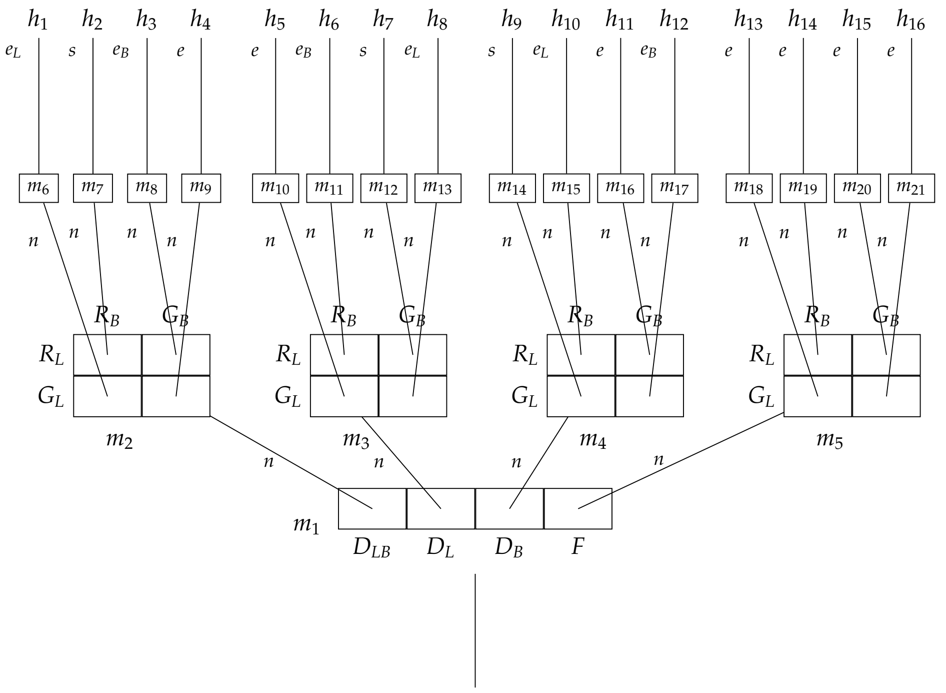

h_1 h_2 h_3 h_4 h_5 h_6 h_7 h_8 h_9 h_{10} h_{11} h_{12} h_{13} h_{14} h_{15} h_{16}
e_L s e_B e e e_B s e_L s e_L e e_B e e e e
m_6 m_7 m_8 m_9 m_{10} m_{11} m_{12} m_{13} m_{14} m_{15} m_{16} m_{17} m_{18} m_{19} m_{20} m_{21}
n
R_B G_B
R_L G_L
m_2 m_3 m_4 m_5
m_1
D_{LB} D_L D_B F

Figure 3.1: *Bomb situation.*

- *Middle level:* at the middle level we find moments m_2–m_5. These moments lie in the possible futures of m_1, and, intuitively speaking, they occur at the same chronological instant. The histories passing through each of these moments are partitioned according to the choices available to *Luther* and *Benji* (after *Ethan* chose and executed one of his available actions at m_1). Since *Ethan* cannot choose anything at the moments of the middle level, $\mathbf{Choice}^{m_i}_{Ethan} = \{H_{m_i}\}$ for all i in 2–5. Frame condition (IA) *independence of agency* ensures that, in Figure 3.1, the layouts of choice-partitions at these moments look like normal-form games. Therefore, *Luther* can be thought of as the row player, and *Benji* as the column player. The actions respectively available to them at all the moments of the middle level are the same:

  – R_L (R_B) stands for 'cutting the red wire of bomb L (B).'

  – G_L (G_B) stands for 'cutting the green wire of bomb L (B).'

- *Top level:* at the top level we find m_6–m_{21}, where the actions executed by *Luther* and *Benji* take effect.

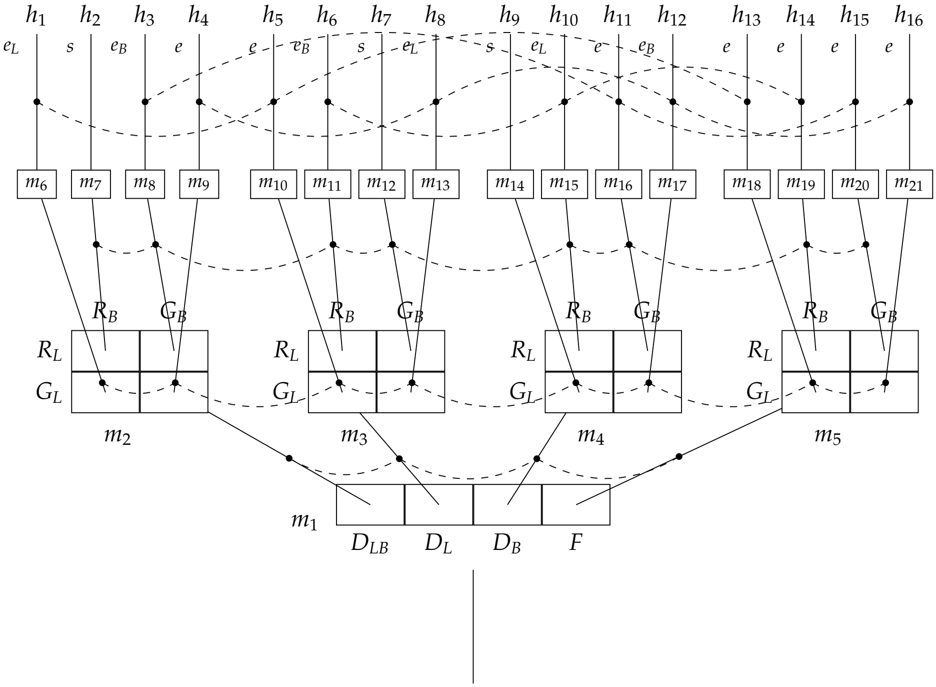
Let me now use formulas of $\mathcal{L}_{\mathsf{KX}}$ to study the bomb situation. In Figure 3.1, e_L, resp. e_B, denotes the atomic proposition 'bomb L, resp. bomb B, explodes'; s stands for 'the bombs are defused and safety is achieved'; e abbreviates $e_L \wedge e_B$; and n abbreviates $\neg e_L \wedge \neg e_B$. Thus, according to Definition 3.3, two examples of the evaluation of formulas are: $\mathcal{M}, \langle m_2, h_1\rangle \models \Box n$ and $\mathcal{M}, \langle m_6, h_1\rangle \models e_L$: *at* $\langle m_2, h_1\rangle$ *it was settled that the bombs have not exploded, but at the next index bomb L has exploded*. Thus, $\mathcal{M}, \langle m_1, h_1\rangle \models \Box n \wedge XXe_L$: *at* $\langle m_1, h_1\rangle$ *it was settled that the bombs have not exploded, but in the next moment of the next moment of* m_1*, bomb L has exploded along history* h_1.

For both cases in Example 3.4, the actual history is h_{10}, where at the bottom level *Ethan* only activated the mechanism of bomb L and at the middle level *Luther* and *Benji* both cut the red wires of their respective bombs. Thus, at $\langle m_4, h_{10}\rangle$ *Luther* has chosen R_L, and *Benji* has chosen R_B, constraining H_{m_4} to the singleton $\{h_{10}\}$. Observe, then, that $\mathcal{M}, \langle m_4, h_{10}\rangle \models n \wedge [Ags]^X e_L$: *at* $\langle m_4, h_{10}\rangle$ *bomb L has not exploded and bomb B has not exploded, but the bomb squad has seen to it that bomb L will explode at next indices—by cutting the wires in a wrong combination*. In contrast, suppose that at m_4 *Luther* chooses G_L, and *Benji* chooses R_B. This means that their actions constrained H_{m_4} to $\{h_9\}$, where $\mathcal{M}, \langle m_4, h_9\rangle \models [Ags]^X s$. *at* $\langle m_4, h_9\rangle$ *the bomb squad has seen to it that both bombs will be defused at next indices—by cutting the wires in a right combination*.

As additional examples to illustrate the evaluation of formulas involving xstit-theoretic operators, consider the following: $\mathcal{M}, \langle m_1, h_{10}\rangle \models X[Luther]^X e_L$: *at* $\langle m_1, h_{10}\rangle$*'s next index* Luther*'s action will cause bomb L to explode (at next indices)*; and $\mathcal{M}, \langle m_3, h_7\rangle \models Y\Diamond X[Luther]^X e_L$: *at* $\langle m_3, h_7\rangle$*'s previous index it was possible that in the next moment* Luther *would have seen to it that bomb L would explode (at next indices).*

This chapter is concerned with the information available to agents and, more specifically, with what this information says about (a) agents' knowledge during the different stages of the decision-making process, and (b) what agents are able to knowingly do. To illustrate instances of these two aspects—in a context of responsibility attribution—let me discuss some examples of the evaluation of epistemic modalities. I focus on *Luther*'s epistemic states, since it is because of them that case a is different from case b in Example 3.4.

In Figure 3.2, resp. Figure 3.3, *Luther*'s epistemic states for Example 3.4 a, resp. Example 3.4 b, are represented with the indistinguishability relation given by dashed lines (where reflexive loops are omitted).



h_1 h_2 h_3 h_4 h_5 h_6 h_7 h_8 h_9 h_{10} h_{11} h_{12} h_{13} h_{14} h_{15} h_{16}
e_L s e_B e e e_B s e_L s e_L e e_B e e e e
m_6 m_7 m_8 m_9 m_{10} m_{11} m_{12} m_{13} m_{14} m_{15} m_{16} m_{17} m_{18} m_{19} m_{20} m_{21}
R_B G_B
R_L G_L
m_2 m_3 m_4 m_5
m_1
D_{LB} D_L D_B F

Figure 3.2: *Example 3.4 a with epistemic state of* Luther.

In both these figures, the valuation of atomic propositions is explicit only at the top level. The reason is that I focus on the outcomes to analyze the alternatives for the epistemic states of *Luther*. Thus, consider the level-by-level comparison between Figure 3.2 and Figure 3.3 below.

- Bottom level: in both figures, at every index based on m_1 *Luther* (and *Benji*, for that matter) did not know what *Ethan* had done. This is represented in the diagram by linking the four clusters of histories passing through m_1 with the dashed lines.

- Middle level: in both figures, at every index based on m_2–m_5 *Luther* was able to distinguish between either cutting the red wire or cutting the green wire of bomb L. Thus, if r_L denotes the proposition 'the red wire of bomb L is cut,' then, for all i in 2–5 and $h \in H_{m_i}$, $\mathcal{M}, \langle m_i, h\rangle \models K_{Luther}[Luther]^X r_L \vee K_{Luther}[Luther]^X \neg r_L$: *at all indices based on moments of the middle level either* Luther *knew that he had cut the red wire of bomb L or he knew that he had not*

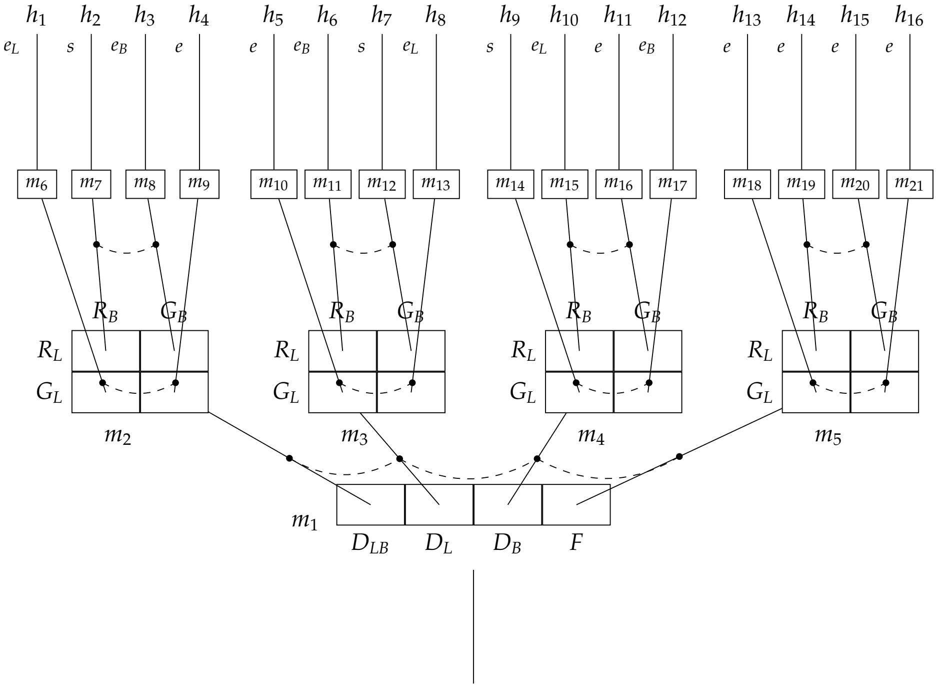


Figure 3.3: *Example 3.4 b with epistemic state of* Luther.

cut the red wire of bomb L (with effects at next indices). Thus, here I assume that *Luther* (and *Benji*) knew which of his available actions was chosen and executed at a given index.[7]

Now, in Figure 3.2 for Example 3.4 a, the actual history h_{10} is such that $\mathcal{M}, \langle m_4, h_{10} \rangle \models [Luther]^X e_L \wedge \Box \neg K_{Luther} [Luther]^X e_L$: *at* $\langle m_4, h_{10} \rangle$ Luther *has seen to it that bomb L will explode at next indices, but it was impossible for* Luther *to have knowingly seen to it that bomb L would explode*. In contrast, in Figure 3.3 for Example 3.4 b, $\mathcal{M}, \langle m_4, h_{10} \rangle \models K_{Luther} [Luther]^X e_L$: *at* $\langle m_4, h_{10} \rangle$ Luther *knew that he has seen to it that bomb L will explode at next indices*.

These are the circumstances that make case a fundamentally different from case b. In case a, although in the actual situation *Luther* made bomb *L* explode by cutting its red wire, his lack of information regarding *Ethan*'s actions implies that there was no way in which he could have known the

[7] This is related to what epistemic game theorists call *knowledge of one's own action*, a property that will be addressed in Section 3.3. Although it holds in this example, not all *ebdt*-frames satisfy it.

consequences of his choice while cutting the red wire. In this sense, one would say that, although he is *causally* responsible for making the bomb explode, he is not *informationally* responsible (see the discussion on Broersen's three categories of responsibility in Chapter 1, p. 5, and Subsection 3.5.2 in this chapter's conclusion). Thus, ultimately he should not be blamed for the unfortunate outcome. In contrast, in case b *Luther* knew what *Ethan* had done, so he was able to distinguish the moment that he and *Benji* found themselves at: in Figure 3.3, at all indices based on moments of the middle level *Luther* was only uncertain about *Benji*'s concurrent choice. In the actual situation, then, *Luther* knowingly set off bomb *L*. In this sense, one would say that *Luther* was both *causally* and *informationally* responsible for making the bomb explode. Thus, ultimately he should be blamed for the undesirable outcome.

- Top level: in Figure 3.2 for case a, at m_6–m_{21}, although the fate of the bombs has been decided, *Luther* can still be uncertain about the exact cause of the explosion. Observe, for instance, that $\mathcal{M}, \langle m_{11}, h_6 \rangle \models \neg K_{Luther} Y[Benji]^X e_B$: *at* $\langle m_{11}, h_6 \rangle$ Luther *did not know that at the previous index* Benji *had seen to it that bomb B would explode.* In contrast, in Figure 3.3 for case b, at all indices based on moments of the top level *Luther* fully knew the state of the world (with respect to the bomb situation), because *Benji*'s choice was disclosed. Thus, *Luther*'s epistemic states at the top level can be seen to illustrate different instances of *ex post* knowledge—as formalized in the next section.

3.3 *Ex Ante*, *Ex interim*, *Ex Post*, and Know-How

In this section I describe several interpretations that stit theorists have given in the past to our four kinds of knowledge, and I compare them with my proposals. As mentioned in the introduction, there are two goals in mind: one is to clarify overlapping intuitions for the work done in *EST*, addressing potential shortcomings of previous analyses; the other is to present new formal characterizations that are more akin to modelling responsibility with stit theory.[8]

As I see it, all the approaches reviewed here—as well as my proposal—intend to model agents' uncertainty in strategic interaction. In particular, one can discern the following levels of uncertainty, in correlation with the four kinds of knowledge studied here: (a) uncertainty about previous actions, correlated with *ex ante*

[8] Recent trends in modelling responsibility by means of *knowingly doing* and of the *epistemic sense of ability* base these two notions on the differential knowledge across the stages of decision making (see, for instance, Abarca & Broersen, 2019; Broersen, 2008a, 2011a; Horty, 2019; Lorini et al., 2014).

knowledge, (b) uncertainty about the effects of one's own actions, correlated with *ex interim* knowledge, (c) uncertainty about other agents' actions, correlated with *ex post* knowledge, and (d) uncertainty about what an agent can knowingly do, correlated with know-how.

3.3.1 *Ex Ante* Knowledge

Ex ante knowledge is commonly thought of as knowledge that an agent has regardless of both its choice of action and the choices of the other agents (Aumann & Dreze, 2008; Lorini et al., 2014). Previous formalizations of the concept in *EST* all try to model this quality, but from somewhat different viewpoints. To illustrate the knowledge that *I* intend to formalize, consider Example 3.4. In case a, at index $\langle m_4, h_{10}\rangle$—the actual situation—neither *Luther* nor *Benji* had *ex ante* knowledge of the fact that *Ethan* had activated the mechanism in the detonator for bomb *B*. In case b, on the contrary, *Luther* did know this *ex ante*. Therefore, if an agent has some certainty about previous actions, it is easier for that agent to discern things *ex ante*.[9]

<u>*Previous versions*</u>

Lorini et al. (2014) presented an epistemic stit logic with three modalities for *ex ante, ex interim,* and *ex post* knowledge: $K_\alpha^{\bullet\circ\circ}\varphi, K_\alpha^{\circ\bullet\circ}\varphi$, and $K_\alpha^{\circ\circ\bullet}\varphi$, respectively.[10] They based the semantics of all three on an indistinguishability relation for agent α's *ex ante* knowledge, given on Kripke structures of possible worlds. The intersection of this *ex ante* relation with the relation for individual, resp. grand-coalition, agency yields their version of *ex interim*, resp. *ex post*, knowledge.[11] Thus, the proposal follows EGT's natural assumption that *ex interim* knowledge refines *ex ante*, and that *ex post* knowledge in turn refines *ex interim*.

The main problems with Lorini et al.'s system will be discussed when dealing with *ex interim* knowledge. However, the fact that the authors did not enforce any connection between *ex ante* knowledge and historical necessity poses an issue for

[9] By saying that it would be 'easier' for an agent to discern things I mean the following: at least intuitively, if at a given index an agent had certainty about previous actions, then for the agent to know a formula *ex ante* the formula would need to hold at *fewer* indices than the amount at which it would need to hold if the agent did not have such a certainty.

[10] As the reader will soon notice, I use different fonts for operator that can be thought of as being analogous across previous approaches, in order to distinguish them. Since the types of knowledge in all these approaches are different from one another, I do this with the hope of avoiding confusion.

[11] Lorini et al.'s models include a set of labels ('action terms' or 'choice names,' in their own words) such that the possible worlds in the domain of their structures are mapped to this set of labels using functions indexed by agents' coalitions. Intuitively, for a given possible world w and a coalition H, the image of this mapping $\mathcal{A}_H(w)$ stands for the action that the coalition H chooses at world w, and all the worlds of the structure that are mapped to the same label under $\mathcal{A}_H$ constitute a cell within a partition that is identified with the partition of H's available choices.

identifying what α knows *ex ante* with knowledge that is present regardless of α's choice and the other agents' choices. In other words, Lorini et al.'s framework admits situations in which at a world agent α knows φ *ex ante*, but a change of action (by either the same agent or one of the others) implies that α no longer knows φ. In Lorini et al.'s logic, $K_\alpha^{\bullet\circ\circ}\varphi \to \Box K_\alpha^{\bullet\circ\circ}\varphi$ is not valid—where I use $\Box\varphi$ to abbreviate Lorini et al.'s $[\emptyset\ \mathbf{stit}]\varphi$. Thus, since the negation of $K_\alpha^{\bullet\circ\circ}\varphi \to \Box K_\alpha^{\bullet\circ\circ}\varphi$ is consistent with formulas $K_\alpha^{\bullet\circ\circ}\varphi$, $\Box(K_\alpha^{\bullet\circ\circ}\varphi \to [\alpha\ \mathbf{stit}]\varphi)$, and $[\alpha\ \mathbf{stit}]\varphi \wedge \neg\Box\varphi$, for instance, then one can build models of Lorini et al.'s logic where (a) $w \models K_\alpha^{\bullet\circ\circ}\varphi$ (α knows φ *ex ante* at world w), and (b) there exists a world v, that does not lie in the same action-cell of α as w, such that $v \not\models [\alpha\ \mathbf{stit}]\varphi$ and $v \not\models K_\alpha^{\bullet\circ\circ}\varphi$. In cases like this, α's *ex ante* knowledge of φ is not independent of α's choice of action.[12]

Horty and Pacuit (2017) addressed a notion of *ex ante* knowledge in a similar way to Lorini et al.'s. The semantics for their *ex ante* modality $\mathcal{K}_\alpha\varphi$ is based on a primitive indistinguishability relation on the set of moments in a branching-time structure. It is only under the light of their full system that I criticize their approach, which will be discussed when dealing with *ex interim* knowledge and know-how.

My version: Let $\mathcal{M}$ be an *ebdt*-model and φ a formula of $\mathcal{L}_{\mathsf{KX}}$. I take agent α's *ex ante* knowledge to be truths about the next moment that, regardless of all agents' current choices, α knows to be independent of said choices. Thus, at index $\langle m, h\rangle$ α had *ex ante* knowledge of φ iff $\mathcal{M}, \langle m, h\rangle \models \Box K_\alpha \Box X\varphi$—that is, iff at $\langle m, h\rangle$ it was settled that α knew that φ would hold at every next index. For instance, consider Example 3.4. If f_B denotes the proposition 'the mechanism of bomb B has been activated,' then $\mathcal{M}, \langle m_4, h_{10}\rangle \models \neg\Box K_{Luther}\Box XY f_B$ for case a (Figure 3.2): *at* $\langle m_4, h_{10}\rangle$ Luther *did not know* ex ante *that, at the index previous to the next, the mechanism of bomb B had been activated*. In turn, $\mathcal{M}, \langle m_4, h_{10}\rangle \models \Box K_{Luther}\Box XY f_B$ for case b (Figure 3.3): *at* $\langle m_4, h_{10}\rangle$ Luther *knew* ex ante *that, at the index previous to the next, the mechanism of bomb B had been activated*.[13]

Two points must be made. First, observe that my *ex ante* knowledge is a single-agent notion, so that the literature on epistemic logic would consider it as individual or private knowledge (Halpern & Fagin, 1989; van Bethem & Sarenac,

[12] Lorini et al. (2014, Remark 2.6, p. 1320) actually stated that the they did not intend for agents to consider all their available choices as epistemically possible in the *ex ante* sense, but this leads precisely to having choice-dependent *ex ante* knowledge.

[13] An important observation to make is that in *ebdt*-models the primitive indistinguishability relation $\sim_\alpha$ characterizes general uncertainty: whatever holds at all epistemically accessible indices is what α knows. I do not subscribe to any game-theoretic interpretation of the primitive modality $K_\alpha\varphi$. In other words, I do not see $K_\alpha\varphi$ as either *ex ante*, *ex interim*, or *ex post* knowledge of φ. Rather, I formalize the levels of uncertainty with a compositional approach, so to speak, where the differences between the stages of information disclosure are embodied by means of different compositions of the modalities for knowledge, agency, and historical necessity.

2004). However, *ex ante* knowledge has been thought of as arising from strictly non-private information. Aumann and Dreze (2008, p. 80), for instance, stated that at the *ex ante* stage of differential information environments no agent should have any private knowledge. Since my version should only be viewed as individual *ex ante* knowledge, I mention that a good candidate for modelling Aumann and Dreze's *ex ante* knowledge of φ could be formula $\Box C \Box X\varphi$, where C denotes the operator for common knowledge of the grand coalition Ags in an extension of the language $\mathcal{L}_{\mathsf{KX}}$ with common-knowledge modalities (see, for instance, Barwise, 1989; Fagin et al., 1995, for a thorough examination of common-knowledge modalities).

Secondly, Duijf (2018, Chapter 3) mentioned that it is important to account for the chronological dimension of *ex ante* knowledge, because the literature usually considers that is based on information that agents had *before* they and the others chose their actions. My proposal does acknowledge the chronological dimension, to some extent. Observe that an agent has *ex ante* knowledge of φ only if φ holds at next indices. Thus, my version concerns information that is available before the meaningful effects of actions take place. To clarify, since in xstit theory the meaningful effects of choices ensue at next indices, then next indices are the relevant points of reference for the formulas that agents know *ex ante*. However, this is more of a conceptual decision with only subtle, terminological consequences in my logic. Indeed, the temporal operators X and Y allow agents to have *ex ante* knowledge about past and current indices as well. For instance, the aforementioned formula $\Box K_{Luther} \Box XY f_B$ is logically equivalent to $\Box K_{Luther} \Box f_B$, and thus it can be seen to refer to a proposition of the current index.[14] In this case, the terminological subtlety refers to the following distinction: for Example 3.4 b, where at $\langle m_4, h_{10}\rangle$ *Luther* has figured out what *Ethan* did in the previous moment, it is not the case that *Luther* knew f_B *ex ante* at the index; rather, *Luther* knew Yf_B *ex ante*.[15]

[14] From the semantics of $X\varphi$ and of $Y\varphi$ one can see that these modalities are inverses of one another, in the sense that formulas $XY\varphi \leftrightarrow \varphi$ and $YX\varphi \leftrightarrow \varphi$ are valid with respect to the class of *ebdt*-models (see Section 3.4).

[15] A reasonable candidate to characterize *ex ante* knowledge could be formula $\Box K_\alpha \Box \varphi$. This formula is not logically equivalent to $\Box K_\alpha \Box X\varphi$ in *EXST*, because this logic does not include restrictions for the persistence of formulas over time. In other words, I do not stipulate whether a formula that holds at a given index should or should not hold as well at a determined number of succeeding indices. For instance, in Example 3.4 a, $\mathcal{M}, \langle m_4, h_{10}\rangle \models \Box K_{Luther} \Box n$ and $\mathcal{M}, \langle m_4, h_{10}\rangle \not\models \Box K_{Luther} \Box Xn$. If one were to characterize α's *ex ante* knowledge with $\Box K_\alpha \Box \varphi$, then at a given index agent α would have known φ *ex ante* iff, regardless of any choice, α knew that φ was currently true independently of any choice. Intuitively, such a formula captures both the assumptions that *ex ante* knowledge must be independent of choices and that it arises before choices are made and executed. However, unless φ is of the form $X\psi$ for some ψ, in the present setting this candidate for knowledge would not (necessarily) regard meaningful facts related to agency of the current moment.

3.3.2 *Ex Interim* Knowledge

It is assumed that at the *ex interim* stage of decision making an agent's available information expands its *ex ante* knowledge by taking into account its own choice, although not yet the other agents'. If an agent knows φ *ex interim*, then the agent is certain about the fact that φ will occur after the performance of its choice of action and regardless of what other agents choose. For instance, in Example 3.4 b, at the actual index *Luther* knew *ex interim* that bomb L would go off after he chose R_L. Intuitively, if an agent has some certainty about the effects of its own actions, then it is easier for that agent to discern things *ex interim*.

Previous versions

For Lorini et al., the information available to an agent *ex interim* is not only independent of the other agents' choices, but also must cause the agent to discern which action it chooses. In other words, in Lorini et al.'s formalism an agent always knows *ex interim* the action that it performs, because the agent can never be uncertain at the *ex interim* stage about the difference between actions that have different labels (see Footnote 11). Such a condition can be syntactically characterized using propositional constants encoding the execution of action labels: if, for action label A, p_α^A denotes the proposition 'the action A is performed by agent α,' then the following formula is valid in Lorini et al.'s logic: $p_\alpha^A \to K_\alpha^{\circ\bullet\circ} p_\alpha^A$. Following Duijf et al. (2021), I refer to this condition as *knowledge of one's own action* (`KOA`).

Although (`KOA`) is in accordance with EGT, I find it a bit constraining. For instance, consider a variation of Example 3.4 for which *Ethan* did not know *ex interim* the difference between detonator L and detonator B but still knowingly activated one of their mechanisms. Using the terminology of Figure 3.1, then, one would say that *Ethan* could not *ex-interim* discern choice D_B from choice D_L. According to Lorini et al., however, *Ethan*'s incapacity to *ex-interim* discern these choices is not allowed.[16] The constraint is all the more problematic because of its consequences for Lorini et al.'s treatment of responsibility attribution. In their formalism, agents will always be morally responsible for performing a given action, so that claiming that they were uncertain about which action they chose is not a valid excuse. Moreover, the constraint implies that *ex ante* certainty of the current moment forces agents to know *ex interim* all the effects of their own actions: formula $(\Box\varphi \to K_\alpha^{\bullet\circ\circ}\varphi) \to ([\alpha\ \mathbf{stit}]\varphi \to K_\alpha^{\circ\bullet\circ}\varphi)$ is valid, and this further

[16] In this chapter I want to have a formal interpretation of uncertainty that admits cases in which agents could not discern which choice they took even after they have already taken it—in agreement with, for example, Duijf (2018) and Broersen (2011a). Observe that this does not imply the collapse of *ex interim* knowledge to *ex ante*. The latter still refines the former, but not in the strict way that Lorini et al.—and both Horty and Pacuit (2017) and Herzig and Troquard (2006), for that matter—proposed.

restricts options of excusability. On a related note, I refer to the validity in *EST* of formula $[\alpha]\varphi \to K_\alpha\varphi$ as property (AEK) or *all-effects knowledge*, by which an agent knows all the effects of its own actions.

When it comes to the relation between *ex interim* knowledge and action labels, Horty and Pacuit's (2017) ideas are similar to Lorini et al.'s. To disambiguate the epistemic sense of ability from its causal sense, Horty and Pacuit based know-how on a novel semantics for *ex interim* knowledge. For the formalization of this *ex interim* knowledge, they extended atemporal *BST* with both syntactic and semantic components. Syntactically, they incorporated modality $[\alpha\ \mathtt{kstit}]\varphi$—meant to express α's *ex interim* knowledge of φ—into a language with modalities $\mathcal{K}_\alpha\varphi$ for α's knowledge, $[\alpha\ \mathtt{stit}]\varphi$ for α's agency, and $\Box\varphi$ for historical necessity. Semantically, they added action labels to *bt*-models, precisely for the evaluation of $[\alpha\ \mathtt{kstit}]\varphi$. In Horty and Pacuit's framework, then, agent α knowingly sees to it that φ iff at all indices that α cannot distinguish from the one of evaluation α's execution of the same action label enforces φ.[17] As mentioned by Broersen and Abarca (2018a) and by Duijf et al. (2021), the use of types brings two limiting constraints:

1. In order for $[\alpha\ \mathtt{kstit}]$ to be an **S5** operator, the primitive indistinguishability relation—underling modality $\mathcal{K}_\alpha\varphi$—must ensue not between indices but between moments. This limits the class of models to those in which knowledge is moment-dependent and agents cannot discern a non-trivial action from another.[18] Actually, this led Horty and Pacuit to identify $\mathcal{K}_\alpha\varphi$ with α's *ex ante* knowledge of φ, but then a shortcoming is that both instances of knowledge—$\mathcal{K}_\alpha\varphi$ and $[\alpha\ \mathtt{kstit}]\varphi$—satisfy the so-called *own action condition* (OAC) (Abarca & Broersen, 2019; Duijf, 2018; Duijf et al., 2021). This condition is semantically stated by the following rule: for each index $\langle m,h\rangle$, $\langle m,h\rangle \sim_\alpha \langle m,h'\rangle$ for every $h' \in \mathbf{Choice}^m_\alpha(h)$. In Horty and Pacuit's formalism, (OAC) implies the validity of formulas $\mathcal{K}_\alpha\varphi \to [\alpha\ \mathtt{stit}]\varphi$ and $[\alpha\ \mathtt{kstit}]\varphi \to [\alpha\ \mathtt{stit}]\varphi$, so that there is no sense whatsoever in which agents can know more than what they bring about.[19]

[17] Formally, $\mathcal{M},\langle m,h\rangle \models [\alpha\ \mathtt{kstit}]\varphi$ iff for all $\langle m',h'\rangle$ such that $\langle m,h\rangle \sim_\alpha \langle m',h'\rangle$, $Exe^{m'}_\alpha(Lbl_\alpha(\langle m,h\rangle)) \subseteq |\varphi|^{m'}$, where Lbl_α is a function that maps an index to the action label of the action token performed by α at that index, $Exe^{m'}_\alpha$ is a partial function that maps types to their corresponding tokens at moment m', and I write $|\varphi|^{m'}$ to refer to the set $\{h' \in H_{m'}; \mathcal{M},\langle m',h'\rangle \models \varphi\}$. This definition is also discussed in Chapter 4 (see Definition 4.16 on p. 159, the discussion following such a definition, and Subsection 4.4.3).

[18] Horty and Pacuit's models satisfy the following constraint: if $\langle m,h\rangle \sim_\alpha \langle m',h'\rangle$, then $\langle m,h_*\rangle \sim_\alpha \langle m',h'_*\rangle$ for every $h_* \in H_m$ and $h'_* \in H_{m'}$. Under reflexivity of $\sim_\alpha$, the constraint is characterized with schema $\mathcal{K}_\alpha\varphi \to \Box\varphi$. Therefore, an agent can only know things that are settled.

[19] To see how this constraint thwarts an analysis of the interplay between knowledge and agency, consider Example 3.4, and assume that one wants to say that at index $\langle m_4,h_{10}\rangle$ *Luther* knew *in some*

2. The semantics for $[\alpha\ \texttt{kstit}]\varphi$ entails that agents cannot have uncertainty about the actions that they perform at the *ex interim* stage, just as in Lorini et al.'s logic. In other words, Horty and Pacuit's *ex interim* knowledge satisfies (KOA): if p^A_α denotes the proposition 'the action A is performed by α,' then formula $p^A_\alpha \rightarrow [\alpha\ \texttt{kstit}]p^A_\alpha$ is valid.

Horty and Pacuit's models meet another important condition that concerns *uniformity*: the same action types must be available at indistinguishable moments.[20] Following Duijf et al. (2021), I refer to this condition as *uniformity of available action types* (UAAT). Initially thought to be a condition that could not be syntactically characterized in Horty and Pacuit's logic without propositional constants (expressing the performance of a certain action type), Duijf et al. (2021) showed that (UAAT) can in fact be characterized with formula $\Diamond[\alpha\ \texttt{kstit}]\varphi \rightarrow K_\alpha\Diamond[\alpha\ \texttt{stit}]\varphi$.[21]

With the same goal as Horty and Pacuit's—giving semantics for know-how—Herzig and Troquard (2006) presented a version of *ex interim* knowledge under the term *dynamic knowledge*. For this dynamic knowledge, the authors extended basic xstit theory's language with modality $[\alpha\ Kstit]\,\varphi$. In contrast to the approaches of Lorini et al. and of Horty and Pacuit, it is for the semantics of $[\alpha\ Kstit]\,\varphi$ that Herzig and Troquard used a primitive indisinguishability relation, and they defined the truth conditions both for agency—$[\alpha\ Stit]\varphi$—and for static knowledge—$\Box\,[\alpha\ Kstit]\,\varphi$—in terms of it. Still, their formalization ends up working in a similar way to Horty and Pacuit's. First, their versions of knowledge—both static and dynamic—satisfy (OAC), so that there is no sense in which agents can know more than what they bring about. Secondly, their treatment is restricted to situations

sense—not in an *ex ante* or *ex interim* sense, though—what *Benji* would choose. Therefore, *Luther* must have somehow distinguished h_9 from h_{12}, and h_{10} from h_{11}. However, in presence of (OAC), this cannot be the case (see Duijf, 2018, Chapter 3, for a more elaborate discussion). An important observation is that properties (KOA), (OAC), and (AEK) are all different from one another, a difference that carries over to all the logics presently reviewed. Property (KOA) (or *knowledge of one's own action*) means that an agent is able to discern which action it is performing, so that—expressed in terms of *EST*—it can be identified with the validity of $p^A_\alpha \rightarrow K_\alpha p^A_\alpha$ (where p^A_α denotes the proposition 'the action A is performed by α'). Property (OAC) (or *own action condition*) means that an agent does not know more than what it brings about, so that it can be identified with the validity of $K_\alpha\varphi \rightarrow [\alpha]\varphi$. Property (AEK) (or *all-effects knowledge*) means that an agent knows the effects of all its actions, so that it can be identified with the validity of $[\alpha]\varphi \rightarrow K_\alpha\varphi$. In all the logics addressed here, (a) (KOA) neither implies (OAC) nor implies (AEK), (b) (OAC) neither implies (KOA) nor implies (AEK), and (c) (AEK) does not imply (OAC) but does imply (KOA).

[20] See Chapter 2's Subsection 2.4.4 (p. 72) for an initial discussion about the concept of uniformity in *EST*. Virtually all the accounts of knowledge that are reviewed in this chapter agree with some version of this condition. That said, although the examples presented by Lorini et al. (2014) all presuppose uniformity, the authors neither explicitly enforced it nor referred to it.

[21] Condition (UAAT), as well as Duijf et al.'s characterization for it in the logic of Horty and Pacuit (2017), are discussed in thoroughly in Chapter 4's Sections 4.3 and 4.4.

where agents cannot be uncertain at the *ex interim* stage about the actions that they choose. Thus, dynamic knowledge satisfies (KOA): if p^A_α denotes the proposition 'the action A is performed by α,' then formula $p^A_\alpha \to [\alpha\ Kstit]p^A_\alpha$ is valid. Lastly, Herzig and Troquard also favored a condition of uniformity, that corresponds to Horty and Pacuit's (UAAT), according to which the same kind of choices must be available at indistinguishable indices.

Duijf (2018, Chapter 3) mentioned an interesting proposal that somewhat resembles Herzig and Troquard's. According to Duijf, a primitive indistinguishability relation that links indices in instantaneous stit theory characterizes a kind of knowledge that one can already call *ex interim*. His rationale is that such knowledge can be seen as (a) private for an agent, and (b) dependent on the agent's choice. These are two qualities of *ex interim* knowledge about which there is no disagreement in the literature, so his approach is substantiated. What distinguishes his interpretation from Herzig and Troquard's—and from Horty and Pacuit's, for that matter—is that Duijf's *ex interim* knowledge is flexible enough to deal both with uncertainty about one's own actions (it is not the case that agents *must* know their actions *ex interim*) and with cases for which agents know more than what they bring about. Thus, instances of both coarse- and fine-grained knowledge are accounted for.[22] However, Duijf's proposal might be a bit too flexible: it allows situations where an agent knows *ex interim* that another agent is bringing about an effect that is not settled. To be more precise, suppose that φ is not settled. If φ is an effect of agent β's action, then α's knowing this is clearly dependent on β's choice. However, denoting Duijf's operator for *ex interim* knowledge by $\mathbb{K}_\alpha$ and the traditional agency operator by $[\alpha]$, formula $\mathbb{K}_\alpha[\beta]\varphi \land \neg\Box\varphi$ is satisfiable in Duijf's logic. It might be surprising that Lorini et al.'s *ex interim* knowledge also does not satisfy (OAC), so a comparable criticism can be advanced. In fact, in Lorini et al.'s formalism agents can have *ex ante* knowledge of what other agents are bringing about, something atypical.

My version: I identify *ex interim* knowledge with Broersen's (2008a) notion of *knowingly doing* (see Chapter 2's Subsection 2.4.4). Let $\mathcal{M}$ be an *ebdt*-model and φ a formula of $\mathcal{L}_{\mathsf{KX}}$. Then at index $\langle m,h\rangle$ agent α had *ex interim* knowledge of φ iff $\mathcal{M},\langle m,h\rangle \models K_\alpha[\alpha]^X\varphi$—that is, iff at $\langle m,h\rangle$ α has knowingly enforced φ (at next indices), regardless of what the other agents did. Intuitively, this captures the idea that whatever was known *ex interim* by α (a) must have been known regardless of other agents' choices, and (b) depended on the epistemic equivalents, across indistinguishable indices, of α's choice. For instance, consider

[22] Duijf did not demand that his models satisfy (OAC). However, and as will be pointed out when addressing his formalization of know-how, Duijf did enforce a constraint corresponding to Horty and Pacuit's (UAAT).

Example 3.4 b (Figure 3.3). Here, $\mathcal{M}, \langle m_4, h_{10}\rangle \models K_{Luther}[Luther]^X e_L$: *at* $\langle m_4, h_{10}\rangle$*—the actual situation—*Luther *has knowingly seen to it that bomb L will go off, so that he knew* ex interim *that bomb L would go off.* In contrast, in case a (Figure 3.2), $\mathcal{M}, \langle m_4, h_{10}\rangle \models \neg K_{Luther}[Luther]^X e_L$: *at* $\langle m_4, h_{10}\rangle$ Luther *has set bomb L off, but not knowingly (i.e., without* ex interim *knowledge about it).*

As for how my *ex interim* knowledge compares with previous proposals, three important remarks must be made:

1. I do not impose any condition on the primitive indistinguishability relation $\sim_\alpha$ that would exclude uncertainty of one's own action in the *ex interim* stage. In other words, my *ex interim* knowledge does not satisfy (`KOA`). Thus, cases of coarse-grained knowledge—with clear implications for responsibility attribution—are allowed: agents can be excused for bringing about an undesirable outcome if they were uncertain about the action that they chose. To illustrate this, consider the variation of Example 3.4 mentioned on p. 92, where *Ethan* could not *ex interim* discern between detonator *L* and detonator *B*. Let $p^{D_B}_{Ethan}$ denote the proposition 'the action D_B is performed by *Ethan*' (i.e., the action of activating the mechanism in the detonator of bomb *B* is performed by *Ethan*), and let $\sim_{Ethan}$ be defined so that, for i, j in 5–12, $\langle m_1, h_i\rangle \sim_{Ethan} \left\langle m_1, h_j\right\rangle$. Then $\mathcal{M}, \langle m_1, h_{10}\rangle \models \neg K_{Ethan}[Ethan]^X Y p^{D_B}_{Ethan}$: *at* $\langle m_1, h_{10}\rangle$ Ethan *did not know* ex interim *that he chose action* D_B.[23] Thus, if *Luther* and *Benji* unfortunately cut both red wires and bomb *L* explodes, the squad can be excused for failing in their mission if *Ethan* claims that it was impossible for him to discern one detonator from the other.

2. As evidenced by the difference between cases a and b of Example 3.4, my *ex interim* knowledge does not satisfy (`AEK`) or *all-effects knowledge*: in case a, although *Luther* saw to it that bomb *L* exploded, he did so without *ex interim* knowledge about it. Furthermore, and in contrast to Lorini et al.'s version, full certainty about the present moment does not imply that agents will know *ex interim* the effects of all their actions. An example of this can be seen in the situation devised for *Ethan* in the above item.

3. My framework is flexible enough so that agents can *in some sense* know more than what they bring about. Since (`OAC`) is not imposed, instances of

[23] Observe that, in the particular case of propositional constants encoding the performance of an action (such as $p^{D_B}_{Ethan}$ above), these constants are interpreted as instantaneous effects of performing the action. To be consistent with my exposition of *ex interim* knowledge, then, I used formula $Yp^{D_B}_{Ethan}$ as an effect (at the next index) of having performed action D_B at the current index (see the discussion of instantaneous effects of actions in Section 3.2, just after Definition 3.3, and also Footnote 15).

finer-grained knowledge are thus accounted for. However, agents cannot *ex-interim* know more than what they bring about: reflexivity of $\sim_\alpha$ entails that $K_\alpha[\alpha]^X\varphi \to [\alpha]^X\varphi$ is valid with respect to the class of *ebdt*-models.[24]

3.3.3 *Ex Post* Knowledge

At the last stage of information disclosure, it is revealed to all agents which actions they chose. Game theorists call the knowledge that arises at this point *ex post* knowledge. Although Aumann and Dreze (2008, p. 80) stated that at the *ex post* stage "all information is revealed to all," I take this to refer to information that results from disclosing *Ags*'s joint choice. Now, EGT also considers that *ex post* knowledge is knowledge that is attained *after* all agents have performed their actions, which adds a chronological dimension to the concept. Intuitively, if an agent has some certainty about the effects of joint action, then it is easier for that agent to know things *ex post*. In Example 3.4 a, for instance, I consider that the fact that *Ags* caused a bomb to go off, and did so unknowingly, is an instance of *ex post* knowledge at $\langle m_4, h_{10}\rangle$. In contrast, I do not consider that it should also be *ex post* knowledge that it was *Luther* who set off the bomb, and that it was bomb L. In fact, although for Example 3.4 b one would definitely say that at $\langle m_4, h_{10}\rangle$ *Luther* knew *ex post* that he set off the bomb, in my opinion *Benji* should not know, even *ex post*, that *Luther* caused the explosion; what he should know *ex post* is that *Luther* chose R_L, and that he himself chose R_B.

<u>*Previous versions*</u>

Out of all the approaches reviewed so far, only Lorini et al. (2014) included some treatment of *ex post* knowledge in *EST*. For them, *ex post* knowledge concerns facts that hold after constraining the possible worlds to those joint actions—of the grand coalition—that are epistemically equivalent to the current joint action of the grand coalition. The relation that provides semantics for $K_\alpha^{\circ\circ\bullet}\varphi$ is built by intersecting the primitive relation of α's *ex ante* knowledge with the relation for group agency of the grand coalition (itself the intersection of the agents' relations for individual agency). A point to be made is that the instantaneous nature of Lorini et al.'s action semantics excludes an analysis of the chronological dimension of *ex post* knowledge. For instance, their framework does not allow that agent α knows *ex post* that it brought about φ without knowing *ex interim* that it brought about φ. In othr words, formula $K_\alpha^{\circ\circ\bullet}[\alpha\ \mathbf{stit}]\varphi \to K_\alpha^{\circ\bullet\circ}[\alpha\ \mathbf{stit}]\varphi$ is valid.

[24] Comparing my *ex interim* knowledge with Duijf's (2018), observe that, with mine, an agent could not have known *ex interim* that another agent would enforce φ (at next indices) without it being settled that φ would hold at next indices. In other words, $K_\alpha[\alpha]^X Y[\beta]^X\varphi \to \Box X\varphi$ is valid with respect to *ebdt*-models. As mentioned before, with Duijf's version, an agent can know *ex interim* that another agent is bringing about an effect that is not settled.

Duijf (2018, Chapter 3), for his part, merely commented on two possibilities for the formalization of *ex post* knowledge. Accounting for its chronological flavor, he mentioned that a good candidate for the individual case is $X\mathbb{K}_\alpha\varphi$, where, as before, I use $\mathbb{K}$ to denote the operator for Duijf's *ex interim* knowledge. Accounting for the public nature of *ex post* knowledge, the other good candidate proposed by Duijf is $D\varphi$, where D is the operator for *distributed knowledge* of the grand coalition.[25] Although both proposals have benefits, Duijf did not explore them further in his approach, which includes neither the 'next' operator nor an operator for distributed knowledge. Now, the first account ($X\mathbb{K}\varphi$) might be a bit too fine-grained: it implies that an agent's *ex interim* knowledge at the next index is the same as its current *ex post* knowledge, so that *ex post* knowledge would take into account an agent's choice at the next index. This is something atypical, to say the least. The only criticism to the second account ($D\varphi$) that I can presently put forward is that it does not consider the chronological dimension of *ex post* knowledge.

My version: I propose a version of individual *ex post* knowledge as follows: let $\mathcal{M}$ be an *ebdt*-model and φ a formula of $\mathcal{L}_{\mathsf{KX}}$. Then at index $\langle m,h\rangle$ agent α had *ex post* knowledge of φ iff $\mathcal{M},\langle m,h\rangle \models X\ K_\alpha\ Y\ [Ags]^X\varphi$—that is, iff at $\langle m,h\rangle$'s next index α comes to know that the choice of the grand coalition enforced φ at next indices. Thus, for an index of evaluation, *ex post* knowledge (a) is attained at the next index, and (b) concerns the interaction that happened at the index of evaluation. One can think of α's *ex post* knowledge as facts supported by those joint actions of the grand coalition that—under α's view—are indistinguishable from the grand coalition's current joint action.[26] In this way, in Example 3.4 a (Figure 3.2) $\mathcal{M},\langle m_4,h_{10}\rangle \models X\ K_{Luther}\ Y\ [Ags]^X(e_L \vee e_B)$: *at* $\langle m_4,h_{10}\rangle$ *the fact that either bomb B or bomb L will explode at next indices is known* ex post *by* Luther. For case b (Figure 3.3), $\mathcal{M},\langle m_4,h_{10}\rangle \models \neg X\ K_{Benji}\ Y\ [Ags]^X(Y[Luther]^X e_L)$: *at* $\langle m_4,h_{10}\rangle$ Benji *did not know* ex post *that* Luther *set off bomb L*. Other interesting cases appear in the non-actual situations where the bombs were defused. In Example 3.4 a, for instance, letting f_B denote the

[25] Distributed knowledge is an instance of group knowledge commonly thought to refer to the knowledge that results from the members' sharing of their information (see, for instance, Fagin et al., 1995; Gerbrandy, 1998; Halpern & Fagin, 1989, for a thorough examination of distributed knowledge).

[26] Observe that I do not interpret α's *ex post* knowledge as facts that the current joint action of the grand coalition would force α to know. In other words, I do not use formula $[Ags]^X K_\alpha\varphi$ to express agent α's *ex post* knowledge at an index. That said, this formula is also a good candidate, since it represents the knowledge that is supported by the actual joint action. The reason for not using it is that I interpret the information that results from the disclosure of the grand coalition's actions in a very particular way: rather than seeing it as the revelation of the actual index's interaction (with the consequent revelation of the actual moment), I see it as those facts for which α knows for certain that they were caused by the interaction of the grand coalition (i.e., by Ags's joint action at the previous index). In other words, *ex post* knowledge concerns not so much those facts that Ags's action forces α to know as those facts about which α knows that they were enforced by Ags.

proposition 'the mechanism of bomb B has been activated,' both $\mathcal{M}, \langle m_4, h_9\rangle \models \neg K_{Luther}(Y[Ethan]^X f_B)$ and $\mathcal{M}, \langle m_4, h_9\rangle \models X\ K_{Luther}\ Y\ [Ags]^X(Y\ Y[Ethan]^X f_B)$. This means that *Luther*—and *Benji*, for that matter—realized at the *ex post* stage that *Ethan* had secured the detonator for bomb B.

Observe that, contrary to Lorini et al.'s formalization, mine admits cases where an agent knows *ex post* that it brought about φ without knowing *ex interim* that it would bring about φ. In other words, $X\ K_\alpha\ Y\ [Ags]^X Y[\alpha]^X\varphi \to K_\alpha[\alpha]^X\varphi$ is not valid with respect to *ebdt*-models. For an example, consider Figure 3.2. Observe that $\mathcal{M}, \langle m_2, h_1\rangle \models X\ K_{Luther}\ Y\ [Ags]^X(Y[Luther]^X e_L)$ and that $\mathcal{M}, \langle m_2, h_1\rangle \not\models K_{Luther}[Luther]^X e_L)$: *at* $\langle m_2, h_1\rangle$ Luther *knew* ex post *that he had set off bomb L, but he did not know this* ex interim.

To deal with the public nature of *ex post* knowledge, one can think of a collective version of the proposal above. Inspired by Duijf's (2018) aforementioned ideas, *ex post* knowledge of φ can be characterized with formula $X\ D\ Y\ [Ags]^X\varphi$, where D is the operator for distributed knowledge of the grand coalition Ags.[27] Intuitively, then, at a given index the grand coalition knew φ *ex post* iff at the next index the grand coalition's sharing of their information supports that they would bring about φ.

Endowed with my three versions of the kinds of knowledge across the stages of decision making, let me address how they interact with each other. Complying with the game-theoretic view, *ex post* refines *ex interim*, which in turn refines *ex ante*:

Proposition 3.5. *Let $\mathcal{M}$ be an* ebdt*-frame. Then for all $\alpha \in Ags$ and all φ of $\mathcal{L}_{\mathsf{KX}}$, the following items hold:*

1. $\mathcal{M} \models \Box K_\alpha \Box X\varphi \to K_\alpha[\alpha]^X\varphi$.
2. $\mathcal{M} \models K_\alpha[\alpha]^X\varphi \to X\ K_\alpha\ Y\ [Ags]^X\varphi$.

Proof. In Section 3.4 I present a proof system Λ_K for *EXST* (Definition 3.6). Both items follow from soundness of Λ_K with respect to *ebdt*-models:

1. Follows from the validity of axiom $\Box\varphi \to \varphi$, of schema $\Box\varphi \to [\alpha]\varphi$, and of schemata ($K$) for $\Box$ and K_α, as well as from preservation of validity under *Modus Ponens* and Necessitation for $\Box$ and K_α.
2. Follows from the validity of schema $[\alpha]\varphi \to [Ags]\varphi$, of axiom $XY\varphi \leftrightarrow \varphi$, of schema $K_\alpha X\varphi \to XK_\alpha\varphi$, and of schema ($K$) for K_α, as well as from preservation of validity under *Modus Ponens* and Necessitation for K_α.

[27] Formally, if one adds $D\varphi$ to $\mathcal{L}_{\mathsf{KX}}$, then, for $\sim_{Ags} := \bigcap_{\alpha\in Ags} \sim_\alpha$, $\mathcal{M}, \langle m, h\rangle \models D\varphi$ iff for all $\langle m', h'\rangle$ such that $\langle m, h\rangle \sim_{Ags} \langle m', h'\rangle$, $\mathcal{M}, \langle m', h'\rangle \models \varphi$.

□

3.3.4 Know-How

When I talk about know-how I refer to the so-called *practical* or *procedural knowledge* of an agent, that takes actions—rather than propositions—as content (see Chapter 2's Footnote 56, p. 72). The intuition is that an agent knows how to do φ iff it has procedural knowledge for bringing about φ. I am not engaging in a circular argument here, because the second statement can be described with precise definitions for *knowledge*, *action*, and *possibility*. Still, I acknowledge that there is a lively debate in the literature as to what being able to bring about φ exactly means, and the question of what it means to be *able in the epistemic sense* adds on further challenges (Ågotnes et al., 2015; Duijf, 2018; Horty & Pacuit, 2017; Naumov & Tao, 2017, 2018; Wang, 2015).

Presently, the concept of know-how is based on *EST*'s approach. Much like Horty and Pacuit (2017), I focus on the differences between what agents can bring about, on the one hand, and what they can knowingly bring about, on the other. Consider the different cases in Example 3.4 (a and b). In case a, one would say that *Luther* and *Benji* did not know how to save the facility. Whether their being able to knowingly save the facility is equal to their knowing how to save it is very much open to debate, but the reader will agree that at least it is necessary for their knowing how. All the authors cited in the previous paragraph agree with this claim, and both Herzig and Troquard and Horty and Pacuit in fact identified knowing how to do φ with the possibility of knowingly bringing about φ. In what follows, I focus my study of the previous literature on individual know-how, leaving its collective counterpart for future work.

Previous versions

Horty and Pacuit (2017) worked on the assumption that an agent is epistemically able to do φ iff it is possible for the agent to have *ex interim* knowledge of φ. Their goal was to formally disambiguate simplified versions of the different cases in my Example 3.4. Using modality $[\alpha\ \texttt{kstit}]\varphi$ (that was reviewed on p. 93), they characterized know-how with formula $\Diamond[\alpha\ \texttt{kstit}]\varphi$. I have already commented on the problems of their system (see also Duijf et al., 2021).

A related approach is Herzig and Troquard's (2006). With their dynamic knowledge modality $[\alpha\ Kstit]\varphi$, the authors proposed that at $\langle m,h\rangle$ agent α knows how to see to it that φ iff $\langle m,h\rangle \models \Box[\alpha\ Kstit]\Diamond[\alpha\ Kstit]X\varphi$—that is, iff $\langle m,h\rangle$ α has static knowledge of the possibility of knowingly bringing about φ. Herzig and Troquard's treatment of know-how is insightful, but their models are constrained by (OAC) and by the condition of uniformity that was mentioned before.

Somewhat related to Herzig and Troquard's version, Duijf (2018, Chapter 3) presented a simple and elegant theory of know-how, that ultimately characterizes individual know-how as follows: at $\langle m, h\rangle$ agent α knows how to see to it that φ iff $\langle m, h\rangle \models \mathbb{K}_\alpha \Diamond \mathbb{K}_\alpha [\alpha]\varphi$—where $\mathbb{K}_\alpha$ is the operator for Duijf's *ex interim* knowledge. As addressed before, Duijf rejects imposing (OAC), so his logic admits finer-grained knowledge. However, a condition known as *uniformity of historical possibility*, which in Duijf's system is syntactically characterized with schema $\Diamond \mathbb{K}_\alpha \varphi \to \mathbb{K}_\alpha \Diamond \varphi$, yields that his formula for know-how is equivalent to $\Diamond \mathbb{K}_\alpha [\alpha]\varphi$.[28]

My version: Let $\mathcal{M}$ be an *ebdt*-model and φ a formula of $\mathcal{L}_{\mathsf{KX}}$. At index $\langle m, h\rangle$ agent α knew how to enforce φ at next indices iff $\mathcal{M}, \langle m, h\rangle \models \Box K_\alpha \Diamond K_\alpha [\alpha]^X \varphi$—that is, iff it was settled that α knew that it was possible for itself to know *ex interim* (or knowingly do) φ.[29] Thus, in Example 3.4 a $\mathcal{M}, \langle m_4, h_{10}\rangle \models \neg \Box K_{Luther} \Diamond K_{Luther} [Luther]^X s$: *at* $\langle m_4, h_{10}\rangle$ Luther *did not know how to defuse the bombs*. In contrast, in case b $\mathcal{M}, \langle m_4, h_{10}\rangle \models \Box K_{Luther} \Diamond K_{Luther} [Luther]^X e_L$: *at* $\langle m_4, h_{10}\rangle$ Luther *did know how to set off a bomb*. Since in both cases *Luther* ultimately does set off a bomb, I consider case a as a situation where he should be excused from moral responsibility of the explosion, while case b is one where he should be held morally responsible for it.[30]

[28] Duijf did not comment on such an equivalence, whose deduction (in Duijf's system) comes from the following argument. Let $(Unif - H)$ denote schema $\Diamond \mathbb{K}_\alpha \varphi \to \mathbb{K}_\alpha \Diamond \varphi$ in Duijf's logic. In light of this schema, one has the following derivation, where 'Subs.' abbreviates 'Substitution':

1. $\vdash \mathbb{K}_\alpha [\alpha]\varphi \to \mathbb{K}_\alpha \mathbb{K}_\alpha [\alpha]\varphi$	Subs. of (4) for $\mathbb{K}_\alpha$
2. $\vdash \Diamond \mathbb{K}_\alpha [\alpha]\varphi \to \Diamond \mathbb{K}_\alpha \mathbb{K}_\alpha [\alpha]\varphi$	1, modal logic
3. $\vdash \Diamond \mathbb{K}_\alpha \mathbb{K}_\alpha [\alpha]\varphi \to \mathbb{K}_\alpha \Diamond \mathbb{K}_\alpha [\alpha]\varphi$	Subs. of $(Unif - H)$
4. $\vdash \Diamond \mathbb{K}_\alpha [\alpha]\varphi \to \mathbb{K}_\alpha \Diamond \mathbb{K}_\alpha [\alpha]\varphi$	2, 3, prop. logic.

The other direction is straightforward from the validity of schema (T) for $\mathbb{K}_\alpha$. A similar derivation can be provided to ensure that Herzig and Troquard's $\Box[\alpha\ Kstit]\Diamond[\alpha\ Kstit]\varphi$ is equivalent to $\Diamond[\alpha\ Kstit]\varphi$.

[29] Observe that modal logic and the validity of formula $XY\varphi \leftrightarrow \varphi$ with respect to *ebdt*-models imply that the proposed formula for know-how—$\Box K_\alpha \Diamond K_\alpha [\alpha]^X \varphi$—is logically equivalent to $\Box K_\alpha \Box XY \Diamond K_\alpha [\alpha]^X \varphi$. Therefore, one can characterize my know-how as follows: agent α knows how to see to it that φ holds at the next moment if α knows *ex ante* that at the moment of performing its actions it is possible for itself to know *ex interim* (or knowingly do) φ.

[30] Although I focus on know-how within *EST*, it is worth discussing some of the approaches in the epistemic extensions of alternating-time temporal logic and of coalition logic (Ågotnes et al., 2015; Hoek & Wooldridge, 2003; Naumov & Tao, 2017, 2018). The reason is that these logics' semantics for know-how are similar to *EST*'s. For instance, Naumov and Tao (2018) and Ågotnes et al. (2015) share many intuitions with Horty and Pacuit (2017). Both approaches characterized know-how as follows: agent α knows how to bring about φ at state s iff there exists a 'strategy' a such that in all states that are epistemically indistinguishable from s in the eyes of α, 'strategy' a will lead to states

3.4 Axiomatization & Logic-Based Properties

It is time to turn our attention to interesting properties of the logic *EXST*. Since they are more akin to the models that appear in the rest of the thesis (as well as to previous approaches' models), in what follows I will work with *ebdt*-models that meet a constraint of uniformity known as *uniformity of historical possibility*. These models are defined just as in Definition 3.2, but they additionally satisfy frame condition ($\mathtt{Unif-H}$): for all $\alpha \in Ags$ and each index $\langle m, h\rangle$, if $\langle m, h\rangle \sim_\alpha \langle m', h'\rangle$, then for every $h_* \in H_m$ there exists $h'_* \in H_{m'}$ such that $\langle m, h_*\rangle \sim_\alpha \langle m', h'_*\rangle$. I refer to the resulting logic as *EXST-u*.[31]

As for the logic-based properties of modalities $\Box\varphi$ and $[\alpha]\varphi$, they are the same as the ones reviewed in Chapter 2's Subsection 2.3.1: both operators are **S5**, and they validate the schemata known as (*SET*) and (*IA*). Operator $[Ags]$ is also **S5**, so that the properties of joint agency of the grand coalition are the same as the ones reviewed for $[\alpha]$. Operator K_α is **S5** as well, so that the properties of knowledge are the ones reviewed in Chapter 2's Subsection 2.4.4: logical omniscience, factivity, positive introspection, and negative introspection. To address the interplay between these modalities and those incorporated from basic xstit theory, I consider it best to first introduce a proof system for *EXST-u*.

3.4.1 A Proof System for *EXST-u*

Definition 3.6 (Proof system for *EXST-u*)**.** *Let Λ_K be the proof system defined by the following axioms and rules of inference:*

at which φ holds. Naumov and Tao's interpretation for the word 'strategy' is different from Ågotnes et al.'s. While the former authors refer to an action label in single-step transitions, the latter use the term as is done in *ATL*, where strategies are functions that assign to each agent and sequence of states a pertinent action. Regardless of the difference, their formalization of know-how depends on the same reasoning: α knows how to do φ iff there exists a *uniform strategy* such that at all epistemically indistinguishable states the transition assigned by the strategy leads to a state at which φ holds. Thus, both accounts involve the idea of uniformity: for Naumov and Tao, their models include a primitive domain of strategies (action labels) that remains uniform for all agents, so that all agents have the same available actions at all states (not only at indistinguishable ones); Ågotnes et al., for their part, admitted the existence of non-uniform strategies and only demanded uniformity as a requirement for know-how. In this respect, it is to Ågotnes et al.'s approach that my formalization is most similar. The two essential differences, then, are that (1) I work within *EST*, where one can model *ex interim* knowledge and base know-how on it, and (2) I only deal with single-step actions taking effect at next indices, leaving the case for long-term strategies/abilities for future work.

[31] Condition ($\mathtt{Unif-H}$) corresponds syntactically to schema $\Diamond K_\alpha\varphi \to K_\alpha\Diamond\varphi$ (see Duijf et al., 2021, for details, as well as Chapter 4's Subsection 4.5). As such, ($\mathtt{Unif-H}$) has two important consequences in *EXST-u*: my formula for *ex ante* knowledge becomes equivalent to $\Box K_\alpha X\varphi$, and my formula for know-how becomes equivalent to $\Diamond K_\alpha[\alpha]^X\varphi$.

- (Axioms) *All classical tautologies from propositional logic; the* **S5** *schemata for* $\Box$, $[\alpha]$, $[Ags]$, *and* K_α*; axiom* (K) *for* X *and* Y*; and the following schemata:*

$$\begin{array}{ll}
YX\varphi \leftrightarrow \varphi & (In1)\\
XY\varphi \leftrightarrow \varphi & (In2)\\
X\varphi \leftrightarrow \neg X\neg\varphi & (DET.S.X)\\
Y\varphi \leftrightarrow \neg Y\neg\varphi & (DET.S.Y)\\
\Box\varphi \rightarrow [\alpha]\varphi & (SET)\\
[\alpha]\varphi \rightarrow [Ags]\varphi & (GA)\\
[Ags]X\varphi \rightarrow [Ags]X\Box\varphi & (NAgs)\\
\textit{For all } m \geq 1 \textit{ and pairwise distinct } \alpha_1, \ldots, \alpha_m, & \\
\bigwedge_{1\leq i\leq m} \Diamond[\alpha_i]\varphi_i \rightarrow \Diamond\left(\bigwedge_{1\leq i\leq m}[\alpha_i]\varphi_i\right) & (IA)\\
K_\alpha X\varphi \rightarrow XK_\alpha\varphi & (NoF)\\
\Diamond K_\alpha p \rightarrow K_\alpha\Diamond p & (Unif-H)
\end{array}$$

- (Rules of inference) *Modus Ponens, Substitution, and Necessitation for the modal operators.*

For $n \in \mathbb{N} - \{0\}$, Λ_{Kn} *is defined as the proof system constructed by adding axiom* $(AgsPC_n)$ *to* Λ_K*, where*

$$\bigwedge_{1\leq k\leq n} \Diamond\left(\left(\bigwedge_{1\leq i\leq k-1} \neg\varphi_i\right) \wedge [Ags]\varphi_k\right) \rightarrow \bigvee_{1\leq k\leq n} \varphi_k \qquad (AgsPC_n).$$

Schemata $(In1)$ and $(In2)$—where 'In' stands for *inverse*—concern the interaction between the 'next moment' operator and the 'last moment' operator. These schemata characterize syntactically that the relations ... *is the successor of...* and ... *is the predecessor of...* behave as inverses of each other, meaning that each point of evaluation is identified with the successor of its predecessor and with the predecessor of its successor (see Footnote 5). Schemata $(DET.S.X)$ and $(DET.S.Y)$—where 'DET.S' stands for *determinicity and seriality*—characterize syntactically that the successor, resp. predecessor, relation is serial and deterministic (also known as functional). Thus, a successor, resp. predecessor, always exists (seriality), and the successor, resp. predecessor, is unique (determinism).

Schemata (SET) and (IA) are standard in *BST*, and they were discussed in Chapter 2's Subsection 2.3.1.

Schema (GA)—where 'GA' stands for *group additiviy*—characterizes syntactically that effects of individual actions are effects of actions of the grand coalition. Axiom (*NAgs*)—where 'NAgs' stands for *no choice for the grand coalition*—characterizes syntactically frame condition *no choice between undivided histories* (with respect to choices of the grand coalition).

Schema (*NoF*)—where 'NoF' stands for *no forget*—characterizes syntactically frame condition *no forget* (`NoF`): if at an index an agent knew that φ would hold at the next index, then at the next index the agent will know that φ holds there.[32] Schema ($Unif-H$)—where 'Unif-H' stands for *uniformity of historical possibility*—characterizes syntactically frame condition *uniformity of historical possibility*: if at an index it was possible for an agent to know φ, then the agent also knew that φ was possible.

Axiom ($AgsPC_n$) is a version of a standard schema in *BST* that characterizes syntactically that the cardinality of the set of available actions of the grand coalition is at most n, at all indices. The reader is referred to Xu's (1994) seminal paper for a discussion of this schema. For elucidation as to how it encodes this cardinality-property, see the proof of Proposition A.15 (p. 120).

Remark 3.7. Some interesting remarks about Λ_{Kn} (with $n \in \mathbb{N} - \{0\}$) are the following:

- Schemata (*SET*) and (*GA*) imply that $\vdash_{\Lambda_{Kn}} \Box\varphi \to [Ags]\varphi$, so that schemata (*NAgs*) and (*T*) for $[Ags]$ imply that $\vdash_{\Lambda_{Kn}} \Box X\varphi \to X\Box\varphi$, which is a theorem that I refer to as (*NX*).
- Necessitation for Y and $\Box$, together with schemata (*In*1) and (*In*2), entails that schema (*NX*) implies that $\vdash_{\Lambda_{Kn}} Y\Box\varphi \to \Box Y\varphi$, a theorem that I refer to as (*NY*).[33]

[32] Observe that frame condition (`NoF`) roughly corresponds to the property of *perfect recall* of knowledge in extensive-form games (see, for instance Bonanno, 2004). In presence of the conditions captured syntactically by schemata (*In*1) and (*In*2), it implies that if an agent knew φ at an index then at successor indices the agent will have known that φ held at the first index mentioned. Whatever an agent knows to have been the case at some past index will at future indices also be known to have been the case (at the mentioned past index). In other words, (`NoF`), (*In*1), and (*In*2) together imply that formula $K_\alpha\varphi \to X^nKY^n\varphi$ is valid, where, for $\Delta \in \{X, Y\}$ and $n \in \mathbb{N} - \{0\}$, $\Delta^n\psi$ denotes the formula that results from applying n-iterations of operator Δ behind φ. In the present theory, (`NoF`) does not imply that an agent knows the past or knows even its own past actions: $Y\varphi \to K_\alpha Y\varphi$, resp. $Y[\alpha]^X\varphi \to K_\alpha Y[\alpha]^X\varphi$, is not valid with respect to *ebdt*-models.

[33] A derivation of (*NY*) is the following, where 'Nec.' abbreviates 'Necessitation' and 'Subs.' abbreviates 'Substitution':

1. $\vdash_{\Lambda_{Kn}} \varphi \to XY\varphi$	(*In*2)
2. $\vdash_{\Lambda_{Kn}} \Box\varphi \to \Box XY\varphi$	1, Nec. and schema (*K*) for $\Box$
3. $\vdash_{\Lambda_{Kn}} Y\Box\varphi \to Y\Box XY\varphi$	2, Nec. and schema (*K*) for Y
4. $\vdash_{\Lambda_{Kn}} Y\Box XY\varphi \to YX\Box Y\varphi$	Nec. and schema (*K*) for Y on subs. of (*NX*)
5. $\vdash_{\Lambda_{Kn}} YX\Box Y\varphi \to \Box Y\varphi$	Subs. of (*In*1)
6. $\vdash_{\Lambda_{Kn}} Y\Box\varphi \to \Box Y\varphi$	3, 4, 5, prop. logic.

- Schemata (*GA*) and (*NAgs*), coupled with schema (4) for $[\alpha]$, imply that, for all $\alpha \in Ags$, $[\alpha]X\varphi \to [\alpha]X\Box\varphi$, which is a theorem that I refer to as (*NA*), characterizing syntactically that agent α cannot make a choice between undivided histories.

- Schema (*GA*) and axiom ($AgsPC_n$) imply that, for all $\alpha \in Ags$,

$$\vdash_{\Lambda_{Kn}} \bigwedge_{1\leq k\leq n} \Diamond\left(\left(\bigwedge_{1\leq i\leq k-1} \neg\varphi_i\right) \wedge [\alpha]\varphi_k\right) \to \bigvee_{1\leq k\leq n} \varphi_k.$$

  I refer to the corresponding theorem schema as (APC_n). This theorem schema is important in the present axiomatization, because it encodes the fact that at each moment each agent will have at most n choices of action to decide from.[34]

- Necessitation for Y and K_α, together with schemata (*In*1) and (*In*2), entails that (*NoF*) implies that $\vdash_{\Lambda_{Kn}} YK_\alpha\varphi \to K_\alpha Y\varphi$.[35]

As for metalogic properties of *EXST-u*, the following result can be shown straightforwardly:

Proposition 3.8 (Soundness of Λ_{Kn}). *For all $n \in \mathbb{N} - \{0\}$, the proof system Λ_{Kn} is sound with respect to the class of* ebdt-*models that additionally satisfy frame condition* $(\mathtt{Unif-H})$.

While showing soundness is straightforward, completeness with respect to *ebdt*-models (with uniformity) is still an open question. However, a completeness result can be given if one takes the discussion to Kripke semantics.

3.4.2 Kripke Semantics for *EXST-u*

Following Broersen (2008a), Lorini and Sartor, and Payette (2014), I will show that the proof system Λ_{Kn} is sound and complete with respect to a class of general multi-modal Kripke-models (see Chapter 2's Subsection 2.3.2) that I refer to as *Kripke*-exs-*models*.

[34] A derivation of (APC_n) can be obtained as follows: using schema (*GA*) and the fact that $\vdash_{\Lambda_{Kn}} (p \to q) \to (\Diamond p \to \Diamond q)$, one can show by induction on n that

$$\vdash_{\Lambda_{Kn}} \bigwedge_{1\leq k\leq n} \Diamond\left(\left(\bigwedge_{1\leq i\leq k-1} \neg\varphi_i\right) \wedge [\alpha]\varphi_k\right) \to \bigwedge_{1\leq k\leq n} \Diamond\left(\left(\bigwedge_{1\leq i\leq k-1} \neg\varphi_i\right) \wedge [Ags]\varphi_k\right).$$

Then (APC_n) follows from ($AgsPC_n$) and propositional logic.

[35] A derivation of this theorem can be obtained by substituting K_α for $\Box$ and (*NoF*) for (*NX*) in the derivation of theorem (*NY*) in Footnote 33.

Definition 3.9 (Kripke-*exs*-frames & models). *A tuple*

$$\left\langle W, Ags, R_\Box, R_X, R_Y, \texttt{Choice}, \{\approx_\alpha\}_{\alpha \in Ags} \right\rangle$$

*is called a Kripke-*exs*-frame iff*

1. *W is a non-empty set of possible worlds and $R_\Box$ is an equivalence relation over W. For $w \in W$, the class of w under $R_\Box$ is denoted by $\overline{w}$.*
2. *R_X and R_Y are relations on W that fulfill the following conditions:*
   - Seriality*: for all $w \in W$, there exist $w', w'' \in W$ such that wR_Xw' and wR_Yw''.*
   - Determinicity (or functionality)*: for all $w \in W$, if wR_Xw_1 and wR_Xw_2, then $w_1 = w_2$; if $wR_Yw'_1$ and $wR_Yw'_2$, then $w'_1 = w'_2$. Thus, the unique element that is R_X-related to w is known as w's* successor *and is denoted by w^{+1}; the unique element that is R_Y-related to w is known as w's* predecessor *and is denoted by w^{-1}.*
   - $(\texttt{Inverse})_{\texttt{K}}$ *$R_X \circ R_Y = Id$ and $R_Y \circ R_X = Id$.*[36]
3. *$\texttt{Choice}$ is a function satisfying the following properties:*
   - *It assigns to each $\alpha \in Ags$ a partition $\texttt{Choice}_\alpha$ of W, given by an equivalence relation R_α. For $w \in W$, the class of w in the partition $\texttt{Choice}_\alpha$ is denoted by $\texttt{Choice}_\alpha(w)$.*
   - *It assigns to the grand coalition Ags a partition $\texttt{Choice}_{Ags}$ of W such that $\texttt{Choice}_{Ags}(v) = \bigcap\limits_{\alpha \in Ags} \texttt{Choice}_\alpha(v)$ for each $v \in W$. The equivalence relation underlying such a partition is denoted by R_{Ags}.*
   - $(\texttt{SET})_{\texttt{K}}$ *For all $w \in W$, $\texttt{Choice}_\alpha(w) \subseteq \overline{w}$ for every $\alpha \in Ags$. This implies that the set $\{\texttt{Choice}_\alpha(v); v \in \overline{w}\}$ is a partition of $\overline{w}$ for every $\alpha \in Ags$, which is denoted by $\texttt{Choice}_\alpha^{\overline{w}}$. Similarly, it implies that $\texttt{Choice}_{Ags}(w) \subseteq \overline{w}$, and that the set $\left\{\texttt{Choice}_{Ags}(v); v \in \overline{w}\right\}$ is a partition of $\overline{w}$, which is denoted by $\texttt{Choice}_{Ags}^{\overline{w}}$.*
   - $(\texttt{IA})_{\texttt{K}}$ *For all $w \in W$, each function $s : Ags \to 2^{\overline{w}}$ that maps α to a member of $\texttt{Choice}_\alpha^{\overline{w}}$ is such that $\bigcap_{\alpha \in Ags} s(\alpha) \neq \emptyset$.*
   - $(\texttt{NAgs})_{\texttt{K}}$ *For all $w \in W$, $R_\Box \circ R_X \circ R_{Ags} \subseteq R_X \circ R_{Ags}$.*[37]

[36] An important note regarding notation: for relations R, S on a given set, I write $R \circ S$ to denote the composition of R and S, such that $x(R \circ S)y$ iff there exists z in the relevant set such that xSz and zRy.

[37] Observe that this condition, together with reflexivity of R_{Ags}, implies the following conditions: $(\texttt{NA})_{\texttt{K}}$ for $\alpha \in Ags$ and $w \in W^{\Lambda_{Kn}}$, $R_\Box \circ R_X \circ R_\alpha \subseteq R_X \circ R_\alpha$; and $(\texttt{NX})_{\texttt{K}}$ $R_\Box \circ R_X \subseteq R_X \circ R_\Box$. In turn, $(\texttt{NX})_{\texttt{K}}$ and $(\texttt{Inverse})_{\texttt{K}}$ imply $(\texttt{NY})_{\texttt{K}} : R_Y \circ R_\Box \subseteq R_\Box \circ R_Y$.

4. *For all $\alpha \in Ags$, $\approx_\alpha$ is an (epistemic) equivalence relation on W. The following conditions must be satisfied:*
   - $(\mathtt{Unif-H})_\mathtt{K}$ *For all $\alpha \in Ags$, if $v, u \in W$ are such that $v \approx_\alpha u$, then for all $v' \in \overline{v}$ there exists $u' \in \overline{u}$ such that $v' \approx_\alpha u'$.*
   - $(\mathtt{NoF})_\mathtt{K}$ *For all $\alpha \in Ags$, $\approx_\alpha \circ R_X \subseteq R_X \circ \approx_\alpha$.*

Frames for which the group-agency condition in item 3 is relaxed to $\mathtt{Choice}_{Ags}(v) \subseteq \bigcap_{\alpha \in Ags} \mathtt{Choice}_\alpha(v)$ *(for all $v \in W$) are called* super-additive frames. *Frames where the cardinalities of partitions* $\mathtt{Choice}^{\overline{w}}_{Ags}$ *and* $\mathtt{Choice}^{\overline{w}}_{\alpha}$ *are at most n, for every $\alpha \in Ags$ and $w \in W$, are called n-*frames.

*A Kripke-*exs*-model $\mathcal{M}$, then, consists of the tuple that results from adding a valuation function $\mathcal{V}$ to a Kripke-*exs*-frame, where $\mathcal{V} : P \to 2^W$ assigns to each atomic proposition a set of worlds (recall that P is the set of propositions in $\mathcal{L}_{KX}$). If one adds a valuation like this to a tuple defining a super-additive frame, then I refer to the model as super-additive. If one adds a valuation like this to a tuple defining an n-frame, then I refer to the model as an n-model.*

Definition 3.10 (Evaluation rules on Kripke models). *Let $\mathcal{M}$ be a Kripke-*exs*-model. The semantics on $\mathcal{M}$ for the formulas of $\mathcal{L}_{KX}$ are defined recursively by the following truth conditions, mirroring Definition 3.3:*

$$\begin{array}{lll}
\mathcal{M}, w \models p & \textit{iff} & w \in \mathcal{V}(p) \\
\mathcal{M}, w \models \neg\varphi & \textit{iff} & \mathcal{M}, w \not\models \varphi \\
\mathcal{M}, w \models \varphi \wedge \psi & \textit{iff} & \mathcal{M}, w \models \varphi \textit{ and } \mathcal{M}, w \models \psi \\
\mathcal{M}, w \models \Box\varphi & \textit{iff} & \textit{for all } w' \textit{ s. t. } wR_\Box w', \mathcal{M}, w' \models \varphi \\
\mathcal{M}, w \models X\varphi & \textit{iff} & \textit{for all } w' \textit{ s. t. } wR_X w', \mathcal{M}, w' \models \varphi \\
\mathcal{M}, w \models Y\varphi & \textit{iff} & \textit{for all } w' \textit{ s. t. } wR_Y w', \mathcal{M}, w' \models \varphi \\
\mathcal{M}, w \models [\alpha]\varphi & \textit{iff} & \textit{for all } w' \textit{ s. t. } wR_\alpha w', \mathcal{M}, w' \models \varphi \\
\mathcal{M}, w \models [Ags]\varphi & \textit{iff} & \textit{for all } w' \textit{ s. t. } wR_{Ags} w', \mathcal{M}, w' \models \varphi \\
\mathcal{M}, w \models K_\alpha\varphi & \textit{iff} & \textit{for all } w' \textit{ s. t. } w \approx_\alpha w', \mathcal{M}, w' \models \varphi.
\end{array}$$

In Chapter 2's Subsection 2.3.2 I mentioned that evaluating formulas of the basic stit-theoretic language and of xstit-theoretic languages on Kripke structures is a standard practice. Actually, Payette (2014) referred to Kripke models for his xstit theory as 'irregular,' using the term 'regular' to refer to branching-time structures that are similar to *bdt*-models (see Chapter 3's Definition 2.19, p. 57).

The main metalogic result for Λ_{Kn}, then, is given by the following theorem.

Theorem 3.11 (Soundness & Completeness of Λ_{Kn}). *For all $n \in \mathbb{N} - \{0\}$, the proof system Λ_{Kn} is sound and complete with respect to the class of Kripke-*exs*-n-models.*

For the proof of this theorem, the reader is referred to Appendix A. Now, proving completeness is one of the technical challenges—and contributions—of this chapter. It involves a long, step-by-step procedure that follows a strategy similar to Lorini and Sartor's (2016). Although I relegate the details to Appendix A, a summary and some discussion seem suitable.

The proof of completeness consists of three steps. This is because the technique of canonical models only yields completeness of Λ_{Kn} with respect to a class of models where R_{Ags} is *included* in $\bigcap_{\alpha \in Ags} R_\alpha$ (i.e., where the other inclusion is not guaranteed). For the sake of illustration, assume that $\mathcal{M}$ is a super-additive model such that R_{Ags} and R_α (where α ranges over Ags) are the only relations defined. There is a well-known method in modal logic (see, for instance Lorini, 2013; Lorini & Sartor, 2016; Schwarzentruber, 2012; Vakarelov, 1992) that can be used to show that, provided that the set of R_{Ags}-equivalence classes is finite (which under the assumption that $R_{Ags} \subseteq \bigcap_{\alpha \in Ags} R_\alpha$ implies that, for all $\alpha \in Ags$, the set of R_α-equivalence classes is also finite), then one can construct a model $\mathcal{M}'$ where $R_{Ags} = \bigcap_{\alpha \in Ags} R_\alpha$ and such that there exists a surjective bounded morphism from $\mathcal{M}'$ to $\mathcal{M}$.[38]

Therefore, completeness of Λ_{Kn} with respect to the class of super-additive models would straightforwardly imply completeness with respect to the class of actual models. However, to adapt this well-known method to the case of *EXST-u* (where the models include many relations other than R_{Ags} and the R_α's), I make use of an intermediate step. Thus, in the first step the standard technique of canonical models is used to prove completeness with respect to Kripke-*exs*-models that are *super-additive*. In the second, intermediate step I prove completeness with respect to super-additive-models where the temporal relations are irreflexive. In the third and last step I use a version of the aforementioned well-known method to prove completeness with respect to the class of actual models.

Now, I mentioned above that completeness with respect to *ebdt*-models (with uniformity) is still an open problem. If $\mathcal{L}_{\mathsf{KX}}$ did not include modality $Y\varphi$, then one might try to use the technique of Schwarzentruber (2012) (see also Canavotto, 2020, Appendix A.1.2) to prove that satisfiability on Kripke-*exs*-*n*-models—of formulas of this restricted language—implies satisfiability on *ebdt*-models.[39] However, this

[38] To successfully apply this particular well-known method, the relations R_α and R_{Ags} *must* be equivalence relations. This is the technical motivation for having a language that also allows formulas expressing instantaneous agency (with semantics based on partitions $\mathbf{Choice}^m_\alpha$ and $\mathbf{Choice}^m_{Ags}$).

[39] Schwarzentruber's (2012) technique can be summarized as follows: for a formula φ of the restricted language and a Kripke structure that satisfies φ at a given world, one first unravels the Kripke structure to obtain another Kripke structure where R_X is irreflexive and such that it satisfies φ as well. In order to use this R_X-irreflexive Kripke structure to build an *ebdt*-model—in the style of Herzig and Schwarzentruber (2008)—one needs to identify worlds with histories (as was addressed in

technique is not successful when formulas of the full $\mathcal{L}_{\mathsf{KX}}$ are considered. One would end up showing that satisfiability on Kripke models implies satisfiability on 'weird' branching-time structures that admit backward branching! That said, there is an unpublished paper by Payette (2012) in which he showed completeness of an xstit logic, whose language does include $Y\varphi$ but includes neither $K_\alpha\varphi$ nor the instantaneous-agency modality $[\alpha]\varphi$, with respect to his regular models, that above were described as very similar to *bdt*-models. The technique used in this unpublished manuscript is complicated, but it might work to prove completeness of Λ_{Kn} with respect to *ebdt*-models. Unfortunately, such an attempt will have to be postponed for future work.

As for the decidability of *EXST-u*, it is definitely a question worth looking into, especially because there has been a recent revival of the interest in the implementation of stit logics in the development and checking of ethical AI (see, for instance, Arkoudas et al., 2005; Calegari et al., 2020; Pereira & Saptawijaya, 2016, see also the thesis's Conclusion, p. 311). It is always risky to venture a conjecture on the decidability of a logic, but I can mention two topic-related results. First, Payette (2014) showed decidability of his xstit logic (the one mentioned in the above paragraph) via finite model property. However, the finite model rendered is only super-additive, meaning that each action available to *Ags* is included in an intersection of individual actions but is not necessarily the same as such an intersection. Secondly, Schwarzentruber (2012) proved that the logic that results from the fragment of $\mathcal{L}_{\mathsf{KX}}$ without $K_\alpha\varphi$ and $Y\varphi$ is decidable. He used a complex method to do so, namely a translation and a truth correspondence between a logic over said fragment and a logic of chains of coalitions.

3.5 Conclusion

This chapter was a stit-theoretic study of the relation between knowledge and agency as components of responsibility. I want to conclude it with a brief discussion of two topics: (a) an extension of stit theory with another—equally important—epistemic component of responsibility: *belief*, and (b) an initial proposal for formalizing the category of informational responsibility (see the discussion on Broersen's three categories of responsibility in Chapter 1, p. 5).

Chapter 2's Subsection 2.3.2). Schwarzentruber achieved a successful identification by first 'correcting' the R_X-irreflexive Kripke structure: if k is the modal depth of X in φ, then one can build yet another Kripke structure satisfying φ at a corresponding world w such that, after k R_X-steps from w, the $R_\Box$-classes of the worlds in this last Kripke structure are singletons. This corrected structure is then used to build an *ebdt*-model satisfying φ.

3.5.1 An Extension with Belief

Suppose that you are a lawyer. You are building the defense for a trial where your client is being charged with negligent homicide. The case is as follows: a patient was admitted to a hospital in urgent need of surgery. The nurses drew up a chart with information for the surgeons, but the figure regarding how long it had been since the patient last ate had a mistake. Anesthetics for this surgery should have been supplied only if the patient had had an empty stomach for at least eight hours, and they were deadly otherwise. Because of the mistake in the chart, the anesthesiologist believed that it was safe to supply the anesthetics, but in fact the patient had had a full meal just one hour before admittance. The anesthetics were supplied, and the patient died. The patient's family sued the anesthesiologist for negligent homicide, and you were hired to defend her. On the causal level, she is responsible for the death. However, as her defense lawyer, you can argue that she should not be held culpable, since she acted upon the false—but justified—belief that the patient had an empty stomach before admittance. Therefore, your job is to prove that the anesthesiologist did not know that the patient had eaten and that she worked under the justified belief that the patient had not eaten.

This example, based on the 1982 film *The Verdict*, reifies the idea that doxastic states should be taken into consideration when deciding whether an agent is culpable for an undesirable outcome. The intuition, then, is that an agent's beliefs amount to reasons for being excused, in those cases where the agent did not meet an obligation and faces a potential punishment. In the example above, the anesthesiologist had a justified belief that the patient had not eaten, so she should be excused for supplying the anesthetics and causing the patient's death.

Although the last two decades have seen a considerable interest in exploring the influence of knowledge on stit-theoretic agency, there are relatively few extensions of *BST* with belief operators (see, for instance, Broersen, 2011c; Wansing, 2006a). Incorporating beliefs into *BST* can be done in many ways, among which I distinguish the following four:

1. Following the tradition of modal epistemic logic (see, for instance, Fagin et al., 1995; Hintikka, 1962; Stalnaker, 2006), one can base a modality $B_\alpha\varphi$—expressing that α believed φ at an index—on a transitive, serial, and euclidean relation on a *bt*-frame. Yielding the typical **KD45** modal logic of belief, this is the road that Wansing (2006a) took in his way to formalizing *doxastic voluntarism* (a philosophical premise by which agents actively decide to acquire certain beliefs).

2. Following the tradition of *dynamic epistemic logic* (*DEL*), that accounts both for belief revision and for the interaction between knowledge and belief (Baltag & Smets, 2006), one can base $B_\alpha\varphi$ on a plausibility preorder on *bt*-frames, so that agent α believes φ iff φ holds at all worlds that α considers plausible enough. This is commonly referred to as a qualitative theory of belief and belief revision.[40]

3. Following decision theory and epistemic game theory (EGT), one can adopt a quantitative interpretation of agentive belief and add probability measures to *bt*-frames, meant to represent subjective degrees of belief for each agent. This is what Broersen (2011c) and Abarca and Broersen (2021a) did, for instance.

4. Related to the previous approach, one can extend the semantics of *BST* with Harsanyi type-spaces (Harsanyi, 1967), so that each moment of a *bt*-frame would be seen as a type-space.

Given the similarities between epistemic stit theory (*EST*) and EGT, a most natural approach to take is the probabilistic one (the third point). According to my joint work with Jan Broersen (Abarca & Broersen, 2021a), factoring a probabilistic semantics of belief into *EST* also paves the way for nuanced formalizations of responsibility, given that it allows to base ideas of belief-driven choice on expected-utility maximization. Since such ideas underlie the responsibility-related extensions of *EST* that are presented in the conclusions of Chapters 4 and 6, I address their basis here.

According to Machina and Viscusi (2013, Chapter 1), the notion that an agent's beliefs are quantifiable by probabilities "is as old as the idea of probability itself, dating back to the second half of the 17$^{\text{th}}$ century." In turn, *Bayesian subjective probability*—meaning the interpretation of probability theory as a theory of individual belief—became a clear trend in formal philosophy with the seminal works of Ramsey (1926) and of Savage (1954). Ever since these works appeared, epistemic game theorists and decision theorists have done much work on quantitative (or graded) models of belief. The idea is that an agent's degrees of belief, with respect to the truth of φ, are represented by probability distributions on the set of states at which φ holds (see, for instance, Aumann, 1999; Baltag & Smets, 2008; Harsanyi, 1967; Pacuit & Roy, 2017).

To transfer this idea to stit theory, one can build probability spaces on a *bt*-frame so that subsets of $I(M \times H)$ (the set of indices) are events to which agents

[40] Enriching *BST* with belief thus allows for a study of the so-called *soft informational attitudes*, opening up possibilities for a theory of belief revision (see, for instance, Baltag & Smets, 2006; Board, 2004) in branching time, an interesting line of research for future work.

assign probability. To simplify both the terminology and the definitions, let me focus on finite discrete structures, where every subset of $I(M \times H)$ is measurable with respect to α's probability function μ_α (for all $\alpha \in Ags$). Thus, for $\alpha \in Ags$ and index $\langle m, h\rangle$, let me first define $\pi_\alpha[\langle m, h\rangle] := \{\langle m', h'\rangle ; \langle m', h'\rangle \sim_\alpha \langle m', h'\rangle\}$ as α's (ex interim*) information set at* $\langle m, h\rangle$. Then one can define $\mu_\alpha : 2^{I(M \times H)} \to [0, 1]$ as a classical discrete probability function (satisfying the *Kolmogorov axioms* (Kolmogorov, 1956)) such that, for each index $\langle m, h\rangle$, $\mu_\alpha(\pi_\alpha[\langle m, h\rangle]) > 0$ (Abarca & Broersen, 2021a).

For $\alpha \in Ags$, function μ_α can then be used to define a semantics for a modality of the form $B_\alpha\varphi$, expressing α's belief of φ. Following Baltag and Smets (2008), here I present the semantics of *probability-1 belief*. The idea is that an agent probability-1 believes φ—written 'p-1 believes φ,' from here on—iff the agent assigns probability 1 to the set of indices where φ is true. Formally, $\mathcal{M}, \langle m, h\rangle \models B_\alpha\varphi$ iff $\mu_\alpha(||\varphi|| \mid \pi_\alpha[\langle m, h\rangle]) = 1$, where $||\varphi||$ denotes the set $\{\langle m, h\rangle \in I(M \times H); \mathcal{M}, \langle m, h\rangle \models \varphi\}$ and $\mu_\alpha(A|B)$ denotes the probability of A conditional on B. Therefore, one says that at index $\langle m, h\rangle$ agent α p-1 believed φ iff α assigned probability 1 to the set of indices at which φ holds, conditional on α's (*ex interim*) information set at $\langle m, h\rangle$.[41]

To illustrate this semantics for p-1 belief, consider the model $\mathcal{M}$ depicted in Figure 3.4, which formalizes the anesthesiologist example mentioned above.

Example 3.12. *Here,* $Ags = \{\text{patient}, \text{doctor}\}$*, and* m_1*–*m_7 *are moments such that* $\sqsubset$ *is defined so as to be represented by the diagram. There are four histories (*h_1*–*h_4*), representing different possibilities for time to evolve according to the actions available both to* patient *and* doctor*. At* m_1 *we find two choices of action available to* patient*:* E_1*, standing for the choice of refusing to eat, and* E_2*, standing for the choice of eating. According to the action chosen by* patient*, the world evolves toward either* m_2 *or* m_3*. At both these moments, it is* doctor*'s turn to act, and her available choices are the following:* L_1 *and* L_3*, standing for supplying anesthetics; and* L_2 *and* L_4*, standing for refusing to supply anesthetics. As implied by the statement of the example,* h_3 *is the actual history.*

The epistemic states that I focus on are those of doctor*, represented with the indistinguishability relation given by dashed lines in Figure 3.4 (where reflexive loops are omitted). Thus, at all indices based on* m_2 *and* m_3 doctor *did not know whether the patient had eaten. The doxastic states of* doctor *are represented by the discrete probability function* $\mu_{doctor} : 2^{I(M \times H)} \to [0, 1]$*, given by the rules:* $\mu_{doctor}\left(\left\langle m_i, h_j\right\rangle\right) = \frac{1}{4}$ *for*

[41] This definition yields a traditional logic of p-1 belief, where B_α is a **KD45** operator and the two following properties hold: *persistence of knowledge* ($\models K_\alpha\varphi \to B_\alpha\varphi$) (see, for instance, van Ditmarsch et al., 2015, Chapter 1) and *full introspection of belief* ($\models B_\alpha\varphi \to K_\alpha B_\alpha\varphi$ and $\models \neg B_\alpha\varphi \to K_\alpha\neg B_\alpha\varphi$) (see, for instance, Baltag & Smets, 2006, 2008; Stalnaker, 2006).

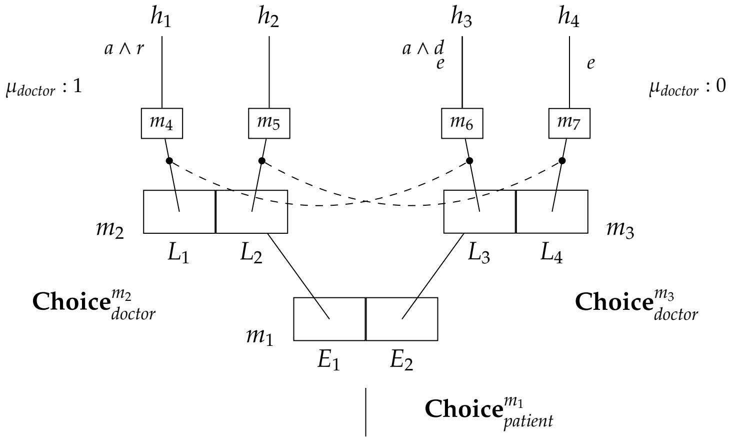


Figure 3.4: *Anesthesiologist example.*

$i, j \in \{1,2\}$, $\mu_{doctor}(\langle m_1, h_i\rangle) = 0$ for $i \in \{3,4\}$, and $\mu_{doctor}(\langle m_3, h_i\rangle) = 0$ for $i \in \{3,4\}$. In the diagram, this is represented by labelling the left-hand side of the model with tag '$\mu_{doctor} : 1$,' and the right-hand side with tag '$\mu_{doctor} : 0$.'

In Figure 3.4, e denotes the atomic proposition 'the patient has eaten,' a stands for 'anesthetics are supplied to the patient,' r stands for 'the patient is ready for surgery,' and d stands for 'the patient will die.' Thus, for all $i \in \{2,3\}$ and $h \in H_{m_i}$, $\mathcal{M}, \langle m_i, h\rangle \models \neg K_{doctor} e \wedge \neg K_{doctor} \neg e$: *at all indices based on* m_2 *and* m_3 doctor *did not know whether the patient had eaten or not*. In turn, $\mathcal{M}, \langle m_3, h_3\rangle \models \neg K_{doctor}[doctor]^X d$: *at the actual index* doctor *did not knowingly kill the patient*.

As for belief, observe that $\mathcal{M}, \langle m_3, h_3\rangle \models B_{doctor}\neg e$: *at the actual index* doctor *p-1 believed that the patient had not eaten*. This is due to the following arguments: *doctor*'s information set at the actual index is $\pi_{doctor}[\langle m_3, h_3\rangle] = \{\langle m_3, h_3\rangle, \langle m_2, h_1\rangle\}$; the set of indices in such an information set at which *patient* has eaten is $\|\neg e\| \cap \pi_{doctor}[\langle m_3, h_3\rangle] = \{\langle m_2, h_1\rangle\}$; at the actual index, then, the probability that *doctor* assigned to the event that the patient had not eaten, conditional on her information set, is $\mu_{doctor}(\|\neg e\| \mid \pi_{doctor}[\langle m_3, h_3\rangle]) = \frac{\mu_{doctor}(\|\neg e\| \cap \pi_{doctor}[\langle m_3, h_3\rangle])}{\mu_{doctor}(\pi_{doctor}[\langle m_3, h_3\rangle])} = \frac{\frac{1}{4}}{\frac{1+0}{4}} = 1$. Thus, *doctor* p-1 believed that the patient had not eaten. Coupled with the facts that *doctor* did not know that the patient had eaten and that *doctor* did not knowingly kill the patient, this p-1 belief provides a good reason for excusing *doctor* from having moral responsibility of the patient's death, despite having caused it.

3.5.2 A Glimpse at Informational Responsibility

All the examples in this chapter convey the intuition that for an agent to be held morally responsible for an undesirable outcome the agent should at the very least have brought it about knowingly. Recall from Chapter 1 (p. 5) that an agent is *causally responsible* for a state of affairs iff the agent is the material author of such a state. In turn, an agent is *informationally responsible* for a state of affairs iff the agent is the material author of such a state and the agent behaved knowingly, or consciously, while bringing it about. Thus, in the bomb example (Example 3.4), *Luther* was causally responsible for detonating a bomb in both case a and case b. However, while in case a he should not be held informationally responsible, in case b he should. In the anesthesiologist example (Example 3.12), *doctor* should not be held informationally responsible for the patient's death, because she did not know that the patient would die after supplying the anesthetics and thus did not knowingly kill the patient. In contrast, in the example of the Clutters' murders the jury dismissed the argument that Hickock and Smith did not act knowingly when they killed the family. Thus, the defendants were indeed informationally responsible for the murders.

Following Lorini et al. (2014), one can think of formula $[\alpha]^X\varphi \wedge \Diamond\neg[\alpha]^X\varphi$, for instance, as a good candidate for syntactically characterizing causal responsibility: agent α was causally responsible for φ iff α saw to it that φ held at next indices and it was possible for α to refrain from seeing to it that φ. In turn, formula $K_\alpha[\alpha]^X\varphi \wedge \Diamond K_\alpha\neg[\alpha]^X\varphi$ is a likewise good candidate for syntactically characterizing informational responsibility: α was informationally responsible for φ iff α knowingly saw to it that φ held at next indices—or knew φ *ex interim* —and it was possible for α to knowingly refrain from seeing to it that φ. Observe, then, that in Example 3.4 a $\mathcal{M}, \langle m_4, h_{10}\rangle \models [Luther]^X e_L \wedge \neg K_{Luther}[Luther]^X e_L$: *at* $\langle m_4, h_{10}\rangle$ Luther *was causally responsible for setting off bomb L, but he was not informationally responsible.* In turn, in Example 3.4 b $\mathcal{M}, \langle m_4, h_{10}\rangle \models K_{Luther}[Luther]^X e_L \wedge \Diamond K_{Luther}\neg[Luther]^X e_L$: *at* $\langle m_4, h_{10}\rangle$ Luther *was informationally responsible for setting off bomb L.*

As a matter of fact, these candidate-formulas greatly influenced Chapter 6's (Subsection 6.3.2) proposal for the syntactic characterization of causal, resp. informational, responsibility, the first two categories in Broersen's classification (p. 5). For the third category—motivational responsibility—it is convenient to first have a formal treatment of the component of responsibility that I have referred to as *intentions*. Chapter 5 is devoted to this respect.

Appendix A Metalogic Results for *EXST-u*

A.1 Soundness

Proposition A.13 (Soundness of Λ_{Kn} (w. r. t. Kripke models))**.** *For all $n \in \mathbb{N} - \{0\}$, the proof system Λ_{Kn} is sound with respect to the class of Kripke-*exs-*n*-*models.*

Proof. Observe that each schema and axiom corresponds to the appropriate relational property in the definition of Kripke-*exs*-*n*-models. Therefore, the proof of soundness is routine. □

A.2 Completeness

Before diving into the actual proof of completeness, here is a succinct description of its structure:

- (Step 1) We build a *canonical structure* from the syntax, where the domain is the set of all Λ_{Kn}*-maximally consistent sets* of formulas of $\mathcal{L}_{\mathsf{KX}}$, and the appropriate relations are defined as is usual in canonical models (see Blackburn et al., 2002, Chapter 4, for details on canonical structures in modal logic). We prove that the canonical structure $\mathcal{M}$ for Λ_{Kn} is a super-additive Kripke-*exs*-*n*-model and show that, for a Λ_{Kn}-consistent formula φ, $\mathcal{M}, w \models \varphi$, where w is the Λ_{Kn}-maximally consistent set including φ. Thus, completeness of Λ_{Kn} with respect to the class of super-additive Kripke-*exs*-*n*-models is shown.

- (Step 2) For any super-additive Kripke-*exs*-*n*-model $\mathcal{M}$, we use a variation of the *unraveling* procedure (see Blackburn et al., 2002, Chapter 2, for details on unraveling) in order to define a structure $\mathcal{M}'$ such that $\mathcal{M}'$ is a super-additive Kripke-*exs*-*n*-model where the 'next' and 'last' relations are irreflexive, and such that there exists a surjective bounded morphism from $\mathcal{M}'$ to $\mathcal{M}$. Using the well-known result of invariance of modal satisfaction under bounded morphisms (Blackburn et al., 2002, Chapter 2), we show completeness of Λ_{Kn} with respect to the class of super-additive Kripke-*exs*-*n*-models where the 'next' and 'last' relations are irreflexive. As mentioned in the body of the chapter, the sole purpose of this intermediate step is to produce a structure that will allow us to adapt the method of Schwarzentruber (2012) (see also Lorini, 2013; Lorini & Sartor, 2016; Vakarelov, 1992) to build an actual model.

- (Step 3) For any super-additive Kripke-*exs*-*n*-model $\mathcal{M}$ where the 'next' and 'last' relations are irreflexive, we use a variation of the mentioned method of

Schwarzentruber (2012) to define a structure $\mathcal{M}'$ such that $\mathcal{M}'$ is a Kripke-*exs*-*n*-model, where $R_{Ags} = \bigcap_{\alpha \in Ags} R_\alpha$ (and where the 'next' and 'last' relations are irreflexive), and such that there exists a surjective bounded morphism from $\mathcal{M}'$ to $\mathcal{M}$. Using the invariance of modal satisfaction under bounded morphisms, we show completeness of Λ_{Kn} with respect to the class of Kripke-*exs*-*n*-models.

Step 1: Canonical Kripke-*exs*-n-Models

In Step 1 we show that the proof system Λ_{Kn} is complete with respect to the class of super-additive Kripke-*exs*-*n*-models. The strategy is to build a canonical structure from the syntax.

Definition A.14 (Canonical Structure)**.** *For $n \in \mathbb{N} - \{0\}$, the tuple*

$$\mathcal{M} = \left\langle W^{\Lambda_{Kn}}, R_\Box, R_X, R_Y, \mathtt{Choice}, \{\approx_\alpha\}_{\alpha \in Ags}, \mathcal{V} \right\rangle$$

is called a canonical structure for Λ_{Kn} *iff*

- *$W^{\Lambda_{Kn}} = \{w; w \text{ is a } \Lambda_{Kn}\text{-MCS}\}$. $R_\Box$ is a relation on $W^{\Lambda_{Kn}}$ defined by the rule: $wR_\Box v$ iff $\Box\varphi \in w \Rightarrow \varphi \in v$ for every φ of $\mathcal{L}_{KX}$. For $w \in W^{\Lambda_{Kn}}$, the set $\left\{v \in W^{\Lambda_{Kn}}; wR_\Box v\right\}$ is denoted by $\overline{w}$. R_X is a relation on $W^{\Lambda_{Kn}}$ defined by the rule: $wR_X v$ iff $X\varphi \in w \Rightarrow \varphi \in v$ for every φ of $\mathcal{L}_{KX}$. R_Y is a relation on $W^{\Lambda_{Kn}}$ defined by the rule: $wR_Y v$ iff for every φ, $Y\varphi \in w \Rightarrow \varphi \in v$.*

- $\mathtt{Choice}$ *is a function that fulfills the following requirements:*

  - *It assigns to each $\alpha \in Ags$ a subset $\mathtt{Choice}_\alpha$ of $2^{W^{\Lambda_{Kn}}}$, defined as follows: let R_α be a relation on $W^{\Lambda_{Kn}}$ such that $wR_\alpha v$ iff $[\alpha]\varphi \in w \Rightarrow \varphi \in v$ for every φ of $\mathcal{L}_{KX}$; if $\mathtt{Choice}_\alpha(v) := \left\{u \in W^{\Lambda_{Kn}}; vR_\alpha u\right\}$, then $\mathtt{Choice}_\alpha := \left\{\mathtt{Choice}_\alpha(v); v \in W^{\Lambda_{Kn}}\right\}$.*

  - *It assigns to the grand coalition Ags a subset $\mathtt{Choice}_{Ags}$ of $2^{W^{\Lambda_{Kn}}}$, defined as follows: let R_{Ags} be a relation on $W^{\Lambda_{Kn}}$ such that $wR_{Ags}v$ iff $[Ags]\varphi \in w \Rightarrow \varphi \in v$ for every φ of $\mathcal{L}_{KX}$; $\mathtt{Choice}_{Ags}$ is then defined analogously to how $\mathtt{Choice}_\alpha$ was defined.*

- *For $\alpha \in Ags$, $\approx_\alpha$ is a relation on $W^{\Lambda_{Kn}}$ given by the rule: $w \approx_\alpha v$ iff $K_\alpha\varphi \in w \Rightarrow \varphi \in v$ for every φ of $\mathcal{L}_{KX}$.*

- *Recall that P is the set of propositions in $\mathcal{L}_{KX}$. Then $\mathcal{V} : P \to 2^{W^{\Lambda_{Kn}}}$ is the canonical valuation, defined so that $w \in \mathcal{V}(p)$ iff $p \in w$.*

Proposition A.15. *For every $n \in \mathbb{N} - \{0\}$, the canonical structure $\mathcal{M}$ for Λ_{Kn} is a super-additive Kripke-*exs*-model where the cardinalities of partitions* $\mathtt{Choice}_{Ags}$ *and* $\mathtt{Choice}_{\alpha}$ *(where α ranges over Ags) are at most n—therefore an n-model.*

Proof. We want to show that $\mathcal{M}$ is a super-additive Kripke-*exs-n*-frame, which amounts to showing that the tuple satisfies the items in the definition of super-additive Kripke-*exs-n*-frames (Definition 3.9).

- Observe that $R_{\Box}$ is an equivalence relation, since Λ_{Kn} includes the **S5** axioms for $\Box$.

- Seriality and determinicity of R_X and R_Y follow from axioms (*DET.S.X*) and (*DET.S.Y*), respectively, by the following arguments. For seriality, take $v \in W^{\Lambda_{Kn}}$. We first show that $z' := \{\psi; Y\psi \in v\}$ and $x' := \{\psi; X\psi \in v\}$ are consistent. Suppose for a contradiction that z' is not consistent. Then there exists a set $\{\psi_1, \dots, \psi_m\}$ of formulas of $\mathcal{L}_{\mathsf{KX}}$ such that $\{\psi_1, \dots, \psi_m\} \subseteq z'$ and (a) $\vdash_{\Lambda_{Kn}} \psi_1 \wedge \dots \wedge \psi_m \rightarrow \bot$. Now, the fact that $\{\psi_1, \dots, \psi_m\} \subseteq z'$ means that $Y\psi_i \in v'$ for every $1 \leq i \leq m$; by Necessitation for Y and its distributivity over conjunction, (a) implies that $\vdash_{\Lambda_{Kn}} Y\psi_1 \wedge \dots \wedge Y\psi_m \rightarrow Y\bot$, which by (*DET.S.Y*) implies that $\vdash_{\Lambda_{Kn}} Y\psi_1 \wedge \dots \wedge Y\psi_m \rightarrow \langle Y \rangle \bot$, but this is a contradiction, since v is a Λ_{Kn}-MCS which includes $Y\psi_1 \wedge \dots \wedge Y\psi_m$. Therefore, z' is consistent. By an analogous argument, x' is consistent as well. Let z be the Λ_{Kn}-MCS that includes z', and let x be the Λ_{Kn}-MCS that includes x' (which exist in virtue of Lindenbaum's Lemma (Blackburn et al., 2002, Chapter 4, p. 199)). Observe that vR_Yz and that vR_Xx. For determinicity, suppose that, besides the existent z and x, there exist z_* and x_* such that vR_Yz_* and vR_Xx_*. We show that $z_* = z$ and that $x_* = x$. For all φ of $\mathcal{L}_{\mathsf{KX}}$, $\varphi \in z_*$ iff $\langle Y \rangle \varphi \in v$ iff (using (*DET.S.Y*)) $Y\varphi \in v$ iff $\varphi \in z$. Therefore, $z_* = z$. An analogous argument, using (*DET.S.X*), shows that $x_* = x$. Axiom (*In*1), resp. (*In*2), ensures that $R_X \circ R_Y = Id$, resp. $R_Y \circ R_X = Id$, according to the following discussion, where we show that $R_X \circ R_Y = Id$ and assume an analogous argument for $R_Y \circ R_X$: take $v \in W^{\Lambda_{Kn}}$, and let z be the unique Λ_{Kn}-MCS such that vR_Yz. We show that zR_Xv: assume that $X\varphi \in z$. Axiom (*DET.S.X*) and the fact that z is a Λ_{Kn}-MCS implies that $X\neg\varphi \notin z$. Suppose for a contradiction that $\varphi \notin v$. Since v is a Λ_{Kn}-MCS, $\neg\varphi \in v$, so that axiom (*In*1) implies that $YX\neg\varphi \in v$. By definition of z, this implies that $X\neg\varphi \in z$, which is a contradiction. Therefore, $\varphi \in v$, and thus zR_Xv.

- Since Λ_{Kn} includes the **S5** schemata for $[\alpha]$, R_α is an equivalence relation for every $\alpha \in Ags$, which implies that $\mathtt{Choice}$, as defined, indeed assigns a partition of $W^{\Lambda_{Kn}}$ to each α. An analogous argument proves that $\mathtt{Choice}$

also assigns a partition of $W^{\Lambda_{Kn}}$ to Ags. Showing that the condition of *superadditivity* holds amounts to proving that $R_{Ags} \subseteq \bigcap_{\alpha \in Ags} R_\alpha$. Schema ($GA$) entails precisely this, as follows: suppose that $wR_{Ags}v$, and assume that $[\alpha]\varphi \in w$ for an arbitrary α. Because of (GA), this implies that $[Ags]\varphi \in w$ as well, so that the supposition that $wR_{Ags}v$ yields that $\varphi \in v$. Therefore, $wR_\alpha v$, but since α was taken arbitrarily, $R_{Ags} \subseteq \bigcap_{\alpha \in Ags} R_\alpha$.

We now verify that Choice satisfies conditions $(\mathtt{SET})_{\mathtt{K}}$ $(\mathtt{IA})_{\mathtt{K}}$, $(\mathtt{NA})_{\mathtt{K}}$, and $(\mathtt{NAgs})_{\mathtt{K}}$:

$(\mathtt{SET})_{\mathtt{K}}$ Since Λ_{Kn} includes $\Box\varphi \to [\alpha]\varphi$ as a schema, then, for all $w \in W^{\Lambda_{Kn}}$, $\mathtt{Choice}_\alpha(w) \subseteq \overline{w}$ for every $\alpha \in Ags$.

$(\mathtt{IA})_{\mathtt{K}}$ We need two intermediate results:

(a) For all $w_* \in W^{\Lambda_{Kn}}$, $w \in \overline{w_*}$ iff $\{\Box\psi; \Box\psi \in w_*\} \subseteq w$. ($\Rightarrow$) Take $w \in \overline{w_*}$ (which means that $w_* R_\Box w$). Take φ of $\mathcal{L}_{\mathsf{KX}}$ such that $\Box\varphi \in w_*$. Since w_* is closed under *Modus Ponens*, axiom (4) for $\Box$ implies that $\Box\Box\varphi \in w_*$. By definition of $R_\Box$, $\Box\varphi \in w$. ($\Leftarrow$) Assume that $\{\Box\psi; \Box\psi \in w_*\} \subseteq w$. Take φ of $\mathcal{L}_{\mathsf{KX}}$ such that $\Box\varphi \in w_*$. By assumption, $\Box\varphi \in w$. Since w is closed under *Modus Ponens*, axiom (T) for $\Box$ implies that $\varphi \in w$. Thus, $w_* R_\Box w$ and $w \in \overline{w_*}$.

(b) If $w_* \in W^{\Lambda_{Kn}}$ and $s : Ags \to 2^{\overline{w_*}}$ maps α to a member of $\mathtt{Choice}_\alpha^{\overline{w_*}}$ such that world $v_{s(\alpha)} \in s(\alpha)$, then $w \in s(\alpha)$ iff $\Delta_{s(\alpha)} = \left\{[\alpha]\psi; [\alpha]\psi \in v_{s(\alpha)}\right\} \subseteq w$. ($\Rightarrow$) Take $w \in s(\alpha)$ (which means that $v_{s(\alpha)} R_\alpha w$). Take φ of $\mathcal{L}_{\mathsf{KX}}$ such that $[\alpha]\varphi \in v_{s(\alpha)}$. Since $v_{s(\alpha)}$ is closed under *Modus Ponens*, schema (4) for $[\alpha]$ implies that $[\alpha][\alpha]\varphi \in v_{s(\alpha)}$. Therefore, by definition of R_α, $[\alpha]\varphi \in w$. ($\Leftarrow$) Assume that $\Delta_{s(\alpha)} = \left\{[\alpha]\psi; [\alpha]\psi \in v_{s(\alpha)}\right\} \subseteq w$. Take φ of $\mathcal{L}_{\mathsf{KX}}$ such that $[\alpha]\varphi \in v_{s(\alpha)}$. By assumption, $[\alpha]\varphi \in w$. Since w is closed under *Modus Ponens*, schema (T) for $[\alpha]$ implies that $\varphi \in w$. Thus, $v_{s(\alpha)} R_\alpha w$ and $w \in s(\alpha)$.

Next, we show that, for all $w_* \in W^{\Lambda_{Kn}}$ and $s : Ags \to 2^{\overline{w_*}}$ just as in item b above, $\bigcup_{\alpha \in Ags} \Delta_{s(\alpha)} \cup \{\Box\psi; \Box\psi \in w_*\}$ is Λ_{Kn}-consistent. First, we show that $\bigcup_{\alpha \in Ags} \Delta_{s(\alpha)}$ is consistent. Suppose that this is not the case. Then there exists a set $\{\varphi_1, \dots, \varphi_n\}$ of formulas of $\mathcal{L}_{\mathsf{KX}}$ such that $[\alpha_i]\varphi_i \in v_{s(\alpha_i)}$ for every $1 \le i \le n$ and

$$\vdash_{\Lambda_{Kn}} [\alpha_1]\varphi_1 \wedge \cdots \wedge [\alpha_n]\varphi_n \to \bot. \tag{3.1}$$

Without loss of generality, assume that $\alpha_i \neq \alpha_j$ for all $j \neq i$ such that $j, i \in \{1, \dots, n\}$—this assumption hinges on the fact that any stit operator distributes over conjunction. Notice that the fact that $[\alpha_i]\varphi_i \in v_{s(\alpha_i)}$ for every $1 \leq i \leq n$ implies that $\Diamond[\alpha_i]\varphi_i \in w_*$ for every $1 \leq i \leq n$. Since w_* is closed under conjunction, $\Diamond[\alpha_1]\varphi_1 \wedge \cdots \wedge \Diamond[\alpha_n]\varphi_n \in w_*$.

Axiom (*IA*) then implies that

$$\vdash_{\Lambda_{Kn}} \Diamond[\alpha_1]\varphi_1 \wedge \cdots \wedge \Diamond[\alpha_n]\varphi_n \to \Diamond\left([\alpha_1]\varphi_1 \wedge \cdots \wedge [\alpha_n]\varphi_n\right). \tag{3.2}$$

Therefore, equations (3.2) and (3.1), imply that

$$\vdash_{\Lambda_{Kn}} \Diamond[\alpha_1]\varphi_1 \wedge \cdots \wedge \Diamond[\alpha_n]\varphi_n \to \Diamond\bot. \tag{3.3}$$

But this is a contradiction, since $\Diamond[\alpha_1]\varphi_1 \wedge \cdots \wedge \Diamond[\alpha_n]\varphi_n \in w_*$, and w_* is a Λ_{Kn}-MCS. Therefore, $\bigcup_{\alpha\in Ags} \Delta_{s(\alpha)}$ is consistent. Secondly, we show that the union $\bigcup_{\alpha\in Ags} \Delta_{s(\alpha)} \cup \{\Box\psi; \Box\psi \in w_*\}$ is also consistent. Suppose that this is not the case. Since $\bigcup_{\alpha\in Ags} \Delta_{s(\alpha)}$ and $\{\Box\psi; \Box\psi \in w_*\}$ are consistent, there must exist sets $\{\varphi_1, \ldots, \varphi_n\}$ and $\{\theta_1, \ldots, \theta_m\}$ of formulas of $\mathcal{L}_{\mathsf{KX}}$ such that $[\alpha_i]\varphi_i \in v_{s(\alpha_i)}$ for every $1 \leq i \leq n$, $\Box\theta_i \in w_*$ for every $1 \leq i \leq m$, and

$$\vdash_{\Lambda_{Kn}} [\alpha_1]\varphi_1 \wedge \cdots \wedge [\alpha_n]\varphi_n \wedge \Box\theta_1 \wedge \cdots \wedge \Box\theta_m \to \bot. \tag{3.4}$$

Let $\theta = \theta_1 \wedge \cdots \wedge \theta_m$. Since $\Box$ distributes over conjunction, $\vdash_{\Lambda_{Kn}} \Box\theta \leftrightarrow \Box\theta_1 \wedge \cdots \wedge \Box\theta_m$, where the fact that w_* is a Λ_{Kn}-MCS closed under logical equivalence implies that $\Box\theta \in w_*$. Thus, (3.4) implies that

$$\vdash_{\Lambda_{Kn}} ([\alpha_1]\varphi_1 \wedge \cdots \wedge [\alpha_n]\varphi_n) \to \neg\Box\theta. \tag{3.5}$$

Once again, assume without loss of generality that $\alpha_i \neq \alpha_j$ for all $j \neq i$ such that $j, i \in \{1, \ldots, n\}$. By an argument analogous to the one used to show that $\bigcup_{\alpha\in Ags} \Delta_{s(\alpha)}$ is consistent, (3.5) implies that

$$\vdash_{\Lambda_{Kn}} \Diamond[\alpha_1]\varphi_1 \wedge \cdots \wedge \Diamond[\alpha_n]\varphi_n \to \Diamond\neg\Box\theta. \tag{3.6}$$

This entails that $\Diamond\neg\Box\theta \in w_*$, but this is a contradiction, since the fact that $\Box\theta \in w_*$ implies with axiom (4) for $\Box$ that $\Box\Box\theta \in w_*$. Now, let u_* be the Λ_{Kn}-MCS that includes $\bigcup_{\alpha\in Ags} \Delta_{s(\alpha)} \cup \{\Box\psi; \Box\psi \in w_*\}$, which exists in virtue of Lindenbaum's Lemma (Blackburn et al., 2002, Chapter 4, p. 199). By intermediate result a, $u_* \in \overline{w_*}$. By intermediate result b, $u_* \in s(\alpha)$ for every $\alpha \in Ags$. Therefore, we have shown that, for all $w_* \in W$, each function $s : Ags \to 2^{\overline{w_*}}$ that maps α to a member of $\mathtt{Choice}^{\overline{w_*}}_{\alpha}$ is such that $\bigcap_{\alpha\in Ags} s(\alpha) \neq \emptyset$, which means that $\mathcal{M}$ satisfies $(\mathtt{IA})_{\mathtt{K}}$.

$(\mathtt{NAgs})_{\mathtt{K}}$ We want to show that $R_\Box \circ R_X \circ R_{Ags} \subseteq R_X \circ R_{Ags}$. Take $v, o \in W^{\Lambda_{Kn}}$ such that $vR_\Box \circ R_X \circ R_{Ags}o$, which means that there exist $v', o' \in W^{\Lambda_{Kn}}$ such that $vR_{Ags}v'$, $v'R_Xo'$, and $o'R_\Box o$. By an argument similar to those in the proof of the second item of the present proposition, we know that $z' := \{\psi; Y\psi \in o\}$ is consistent. Let z be the Λ_{Kn}-MCS that includes z', which exists in virtue of Lindenbaum's Lemma (Blackburn et al., 2002, Chapter 4, p. 199). As shown in the second item of the present proposition, it is the case that zR_Xo. Let us show that $vR_{Ags}z$. Take $[Ags]\varphi \in v$. Axiom ($In2$), schema (K) for $[Ags]$, and Necessitation for $[Ags]$ imply that $[Ags]XY\varphi \in v$. Schema ($NAgs$) then entails that $[Ags]X\Box Y\varphi \in v$, and the assumption that $vR_{Ags}v'$ then yields that $X\Box Y\varphi \in v'$. The assumption that $v'R_Xo'$ implies that $\Box Y\varphi \in o'$, so that the assumption that $o'R_\Box o$ gives that $Y\varphi \in o$, which by construction of z implies that $\varphi \in z$. Therefore, $vR_{Ags}z$, which with the previously shown fact that zR_Xo implies that $vR_X \circ R_{Ags}o$. Thus, $R_\Box \circ R_X \circ R_{Ags} \subseteq R_X \circ R_{Ags}$.

Finally, we show that $card\left(\mathtt{Choice}_{Ags}\right) \leq n$. Take $w \in W^{\Lambda_{Kn}}$. Suppose for a contradiction that $card\left(\mathtt{Choice}_{Ags}\right) > n$. Take pairwise different $c_0, \ldots, c_n \in \mathtt{Choice}_{Ags}$, and take $w_i \in c_i$ for each $0 \leq i \leq n$. Lemma A.16 5 implies that, for all $1 \leq i \leq n$, there exists φ_i of $\mathcal{L}_{\mathsf{KX}}$ such that $[Ags]\varphi_i \in w_i$ and $\varphi_i \notin w_j$ for every $0 \leq j \leq n$ such that $j \neq i$. On the one hand, this implies that ($\star$) $\bigwedge\limits_{1\leq i\leq n} \neg\varphi_i \in w_0$. On the other, it implies that $[Ags]\varphi_1 \in w_1$, $(\neg\varphi_1 \wedge [Ags]\varphi_2) \in w_2, \ldots, (\neg\varphi_1 \wedge \cdots \wedge \neg\varphi_{n-1} \wedge [Ags]\varphi_n) \in w_n$. Thus, one has that $\bigwedge\limits_{1\leq k\leq n} \Diamond\left(\left(\bigwedge\limits_{1\leq i\leq k-1} \neg\varphi_i\right) \wedge [Ags]\varphi_k\right) \in w_0$, which by ($AgsPC_n$) implies that $\bigvee\limits_{1\leq k\leq n} \varphi_k \in w_0$, but this is a contradiction to ($\star$).[42]

- Since Λ_{Kn} includes the **S5** schemata for K_α, $\approx_\alpha$ is an equivalence relation for every $\alpha \in Ags$. We verify that conditions $(\mathtt{Unif-H})_{\mathtt{K}}$ and $(\mathtt{NoF})_{\mathtt{K}}$ are satisfied:

  $(\mathtt{Unif-H})_{\mathtt{K}}$ Take $\alpha \in Ags$, and let $v, u \in W^{\Lambda_{Kn}}$ be such that $v \approx_\alpha u$. Take $v' \in \overline{v}$. We want to show that there exists $u' \in \overline{u}$ such that $v' \approx_\alpha u'$. We show that $u'' = \{\psi; K_\alpha\psi \in v'\} \cup \{\Box\psi; \Box\psi \in u\}$ is consistent. To do so, we first show that $\{\psi; K_\alpha\psi \in v'\}$ is consistent. Suppose for a contradiction that it is not consistent. Then there exists a set $\{\psi_1, \ldots, \psi_n\}$ of formulas of $\mathcal{L}_{\mathsf{KX}}$ such that $K_\alpha\psi_i \in v'$ for every $1 \leq i \leq n$ and (a) $\vdash_{\Lambda_{Kn}} \psi_1 \wedge \cdots \wedge \psi_n \to \bot$. By Necessitation for K_α and its distributivity over conjunction, (a) implies that $\vdash_{\Lambda_{Kn}} K_\alpha\psi_1 \wedge \cdots \wedge K_\alpha\psi_n \to K_\alpha\bot$, but this is a contradiction, since v' is a Λ_{Kn}-MCS that includes $K_\alpha\psi_1 \wedge \cdots \wedge K_\alpha\psi_n$. Next, we show that $u'' =$

[42] Observe that we can use the same argument to show that theorem (APC_n) implies that $card\,(\mathtt{Choice}_\alpha) \leq n$ for every $\alpha \in Ags$.

$\{\psi; K_\alpha\psi \in v'\} \cup \{\Box\psi; \Box\psi \in u\}$ is also consistent. Suppose for a contradiction that it is not consistent. Since $\{\psi; K_\alpha\psi \in v'\}$ and $\{\Box\psi; \Box\psi \in u\}$ are consistent, there must exist sets $\{\varphi_1, \ldots, \varphi_n\}$ and $\{\theta_1, \ldots, \theta_m\}$ of formulas of $\mathcal{L}_{\mathsf{KX}}$ such that $K_\alpha\varphi_i \in v'$ for every $1 \leq i \leq n$, $\Box\theta_i \in w_2$ for every $1 \leq i \leq m$, and (b) $\vdash_{\Lambda_{Kn}} \varphi_1 \wedge \cdots \wedge \varphi_n \wedge \Box\theta_1 \wedge \cdots \wedge \Box\theta_m \to \bot$. Let $\theta = \theta_1 \wedge \cdots \wedge \theta_m$ and $\varphi = \varphi_1 \wedge \cdots \wedge \varphi_n$. Since $\Box$ distributes over conjunction, $\vdash_{\Lambda_{Kn}} \Box\theta \leftrightarrow \Box\theta_1 \wedge \cdots \wedge \Box\theta_m$. Since u is a Λ_{Kn}-MCS closed under logical equivalence, then $\Box\theta \in u$, which by schema (4) for $\Box$ and closure of u under *Modus Ponens* implies that ($\star$) $\Box\Box\theta \in u$ as well. Now, (b) implies that $\vdash_{\Lambda_{Kn}} \varphi \to \neg\Box\theta$ and thus that (c) $\vdash_{\Lambda_{Kn}} \Diamond\varphi \to \Diamond\neg\Box\theta$. Notice that the facts that $K_\alpha\varphi_i \in v'$ for every $1 \leq i \leq n$, that K_α distributes over conjunction, and that v' is a Λ_{Kn}-MCS imply that $K_\alpha\varphi \in v'$. The fact that $v' \in \overline{v}$ implies that $\Diamond K_\alpha\varphi \in v$, so that $(Unif-H)$ entails that $K_\alpha\Diamond\varphi \in v$. Now, this last inclusion implies, with our assumption that $v \approx_\alpha u$, that $\Diamond\varphi \in u$, which by (c) in turn yields that $\Diamond\neg\Box\theta \in u$, contradicting ($\star$). Therefore, u'' is consistent. Let u' be the Λ_{Kn}-MCS that includes u'', which exists in virtue of Lindenbaum's Lemma (Blackburn et al., 2002, Chapter 4, p. 199). By construction, $u' \in \overline{u}$ (in virtue of intermediate result a in the third item), and $v' \approx_\alpha u'$. With this, we have shown that $\mathcal{M}$ satisfies condition $(\mathtt{Unif-H})_{\mathtt{K}}$.

$(\mathtt{NoF})_{\mathtt{K}}$ Take $\alpha \in Ags$. We want to show that $\approx_\alpha \circ R_X \subseteq R_X \circ \approx_\alpha$. Let $w, v \in W^{\Lambda_{Kn}}$ be such that $w \approx_\alpha \circ R_X v$ via y. By similar arguments to the ones used in the second item of this proof (p. 117), we know that $z' := \{\psi; Y\psi \in v\}$ is consistent and that, if z denotes the Λ_{Kn}-MCS that includes z', then $zR_X v$. What remains to be shown is that $w \approx_\alpha z$, so assume that $K_\alpha\varphi \in w$. Axiom ($In2$), schema (K) for K_α, and Necessitation for K_α imply that $K_\alpha XY\varphi \in w$. Schema (NoF) then implies that $XK_\alpha Y\varphi \in w$. By construction, this implies that $K_\alpha Y\varphi \in y$, which in turn implies that $Y\varphi \in v$ and thus that $\varphi \in z$. Therefore, z is such that $w \approx_\alpha z$ and $zR_X v$, which means that $wR_X \circ \approx_\alpha v$. Thus, $\approx_\alpha \circ R_X \subseteq R_X \circ \approx_\alpha$.

□

As is usual with canonical structures, our objective is to prove the so-called *truth lemma*, which says that, for all φ of $\mathcal{L}_{\mathsf{KX}}$ and $w \in W^{\Lambda_{Kn}}$, $\mathcal{M}, w \models \varphi$ iff $\varphi \in w$. This is done by induction on the complexity of φ, and the inductive steps in the cases of each modal operator require previous results (such as the important *existence* lemmas). These existence results are given in Lemma A.16 below.

Lemma A.16 (Existence). *Let $\mathcal{M}$ be the canonical structure for Λ_{Kn}. For every $w \in W^{\Lambda_{Kn}}$ and every φ of $\mathcal{L}_{KX}$, the following items hold:*

1. *$X\varphi \in w$ iff $\varphi \in v$ for every v such that $wR_X v$.*

2. *$Y\varphi \in w$ iff $\varphi \in v$ for every v such that $wR_Y v$.*

3. *$\Box\varphi \in w$ iff $\varphi \in v$ for every $v \in \overline{w}$.*

4. *$[\alpha]\varphi \in w$ iff $\varphi \in v$ for every v such that $wR_\alpha v$.*

5. *$[Ags]\varphi \in w$ iff $\varphi \in v$ for every v such that $wR_{Ags}v$.*

6. *$K_\alpha\varphi \in w$ iff $\varphi \in v$ for every v such that $w \approx_\alpha v$.*

Proof. Take $w \in W^{\Lambda_{Kn}}$ and φ of $\mathcal{L}_{\mathsf{KX}}$. All items are shown in the same way. Take $\triangle \in \{X, Y, \Box, [\alpha], [Ags], K_\alpha\}$, and let $R_\triangle$ stand for the relation upon which the semantics of $\triangle\varphi$ is defined. We show that $\triangle\varphi \in w$ iff $\varphi \in v$ for every $v \in W^{\Lambda_{Kn}}$ such that $wR_\triangle v$.

($\Rightarrow$) Assume that $\triangle\varphi \in w$. Let $v \in W^{\Lambda_{Kn}}$ be such that $wR_\triangle v$. The definition of $R_\triangle$ implies that $\varphi \in v$.

($\Leftarrow$) We work by contraposition. Assume that $\triangle\varphi \notin w$. We show that there is a world v in $W^{\Lambda_{Kn}}$ such that $wR_\triangle v$ and such that φ does not lie within v. For this, let $v' = \{\psi; \triangle\psi \in w\}$, which is shown to be consistent as follows: suppose for a contradiction that v' is not consistent. Then there exists a set $\{\psi_1, \dots, \psi_n\}$ of formulas of $\mathcal{L}_{\mathsf{KX}}$ such that $\{\psi_1, \dots, \psi_n\} \subseteq v'$ and (a) $\vdash_{\Lambda_{Kn}} \psi_1 \wedge \cdots \wedge \psi_n \to \bot$. Now, the fact that $\{\psi_1, \dots, \psi_n\} \subseteq v'$ means that $\triangle\psi_i \in w$ for every $1 \leq i \leq n$. Necessitation for $\triangle$ and its distributivity over conjunction yield that (a) implies that $\vdash_{\Lambda_{Kn}} \triangle\psi_1 \wedge \cdots \wedge \triangle\psi_n \to \triangle\bot$, but this is a contradiction, since w is a Λ_{Kn}-MCS that includes $\triangle\psi_1 \wedge \cdots \wedge \triangle\psi_n$. Now, we define $v' := v' \cup \{\neg\varphi\}$, and we show that it is also consistent. Suppose for a contradiction that it is not consistent. Since v' is consistent, then there exists a set $\{\psi_1, \dots, \psi_n\}$ of formulas of $\mathcal{L}_{\mathsf{KX}}$ such that $\{\psi_1, \dots, \psi_n\} \subseteq v'$ and $\vdash_{\Lambda_{Kn}} \psi_1 \wedge \cdots \wedge \psi_n \wedge \neg\varphi \to \bot$, which then implies that (b) $\vdash_{\Lambda_{Kn}} \psi_1 \wedge \cdots \wedge \psi_n \to \varphi$; By Necessitation for $\triangle$ and its distributivity over conjunction, (b) implies that $\vdash_{\Lambda_{Kn}} \triangle\psi_1 \wedge \cdots \wedge \triangle\psi_n \to \triangle\varphi$. Since w is a Λ_{Kn}-MCS closed under conjunction, then $\triangle\psi_1 \wedge \cdots \wedge \triangle\psi_n \in w$, so that (b) and closure of w under *Modus Ponens* entail that $\triangle\varphi \in w$, contradicting the initial assumption that $\triangle\varphi \notin w$. Let v be the Λ_{Kn}-MCS that includes v'', which exists in virtue of Lindenbaum's Lemma (Blackburn et al., 2002, Chapter 4, p. 199). By construction, $\varphi \notin v$ and $wR_\triangle v$, by definition of $R_\triangle$.[43]

□

Lemma A.17 (Truth Lemma)**.** *Let $\mathcal{M}$ be the canonical structure for Λ_{Kn}. For all φ of $\mathcal{L}_{KX}$ and $w \in W^{\Lambda_{Kn}}$, $\mathcal{M}, w \models \varphi$* iff *$\varphi \in w$.*

[43] Observe that in the cases of $\triangle \in \{[\alpha], [Ags]\}$, schemata (*SET*) and (*GA*) render that the found v actually lies within $\overline{w}$ (if $\Box\theta \in w$, then $[\alpha]\theta \in w$ and $[Ags]\theta \in w$).

Proof. We proceed by induction on the complexity of φ. The cases of propositional letters and of Boolean connectives are standard. For the cases of modal operators, both directions follow from Lemma A.16. □

Proposition A.18 (Completeness w.r.t. super-additive n-models)**.** *The proof system* Λ_{Kn} *is complete with respect to the class of super-additive Kripke-*exs*-*n*-models.*

Proof. Let φ be a Λ_{Kn}-consistent formula of $\mathcal{L}_{\mathsf{KX}}$. Let w be the Λ_{Kn}-MCS including φ, which exists in virtue of Lindenbaum's Lemma (Blackburn et al., 2002, Chapter 4, p. 199). Then the canonical structure $\mathcal{M}$ for Λ_{Kn} is a super-additive Kripke-*exs*-n-model such that $\mathcal{M}, w \models \varphi$, according to Lemma A.17 above. □

Step 2: Irreflexive Super-Additive n-Models

In the next step of our proof of completeness, for a Λ_{Kn}-consistent formula φ of $\mathcal{L}_{\mathsf{KX}}$, we build a super-additive n-model that satisfies φ where the transitive closures of the 'next' and 'last' relations are irreflexive.

The idea is to adapt the unraveling argument to the class of models characterized by the canonical model. If the only relation were R_X, the unraveling would be a straightforward procedure (see Blackburn et al., 2002, Chapter 2, for details on unraveling). However, our models also include R_Y, which is serial and deterministic. If we were to unravel traditionally, the root would not have a predecessor, and then R_Y would not be serial. Therefore, we employ the following strategy: to take each point of the original model and consider it as a 'double root,' that 'grows' with R_X in one direction and 'grows' with R_Y in the opposite direction. The resulting models are forests of bi-directional trees, with one bi-directional tree per world of the original model. These models do not have roots in the traditional sense of the word (where the term 'root' is interpreted as a least element in the strict partial order that gives rise to a tree.) Rather, each tree has a 'center,' given by the sequence of shortest length—length 1—in the tree. In these models, the 'next' and 'last' relations abide by the proper conditions of seriality, determinicity, $(\mathtt{Inverse})_{\mathsf{K}}$, and $(\mathtt{NAgs})_{\mathsf{K}}$.[44]

Before defining this 'double unraveling' formally, we introduce some auxiliary terminology:

- For $w \in W$, $h[w] := \left\{v \in W; wR^*_Y v \text{ or } v = w \text{ or } wR^*_X v\right\}$, where R^*_Y, resp. R^*_X, is the transitive closure of R_Y, resp. R_X. Observe that, for all $w \in W$, R^*_Y, resp. R^*_X, restricted to $h[w]$ is a linear order on $h[w]$.

[44] The two directions of the growing trees are respectively marked by the labels 0 and 1 in the tuples that constitute the model, preventing reflexive loops to arise.

- For $w \in W$ and $i \in \mathbb{N}$, w^{-j} denotes the unique element in $h[w]$ such that $w = w^{-0} R_Y w^{-1} R_Y w^{-2} R_Y \dots w^{-(j-1)} R_Y w^{-j}$. Similarly, w^{+j} denotes the unique element in $h[w]$ such that $w = w^{+0} R_X w^{+1} R_X \dots w^{+(j-1)} R_X w^{+j}$.

Definition A.19 (An unraveling). *Let* $\mathcal{M} = \left\langle W, R_\Box, R_X, R_Y, \texttt{Choice}, \{\approx_\alpha\}_{\alpha \in Ags}, \mathcal{V} \right\rangle$ *be a super-additive Kripke-*exs-*n-model, where equivalence relations* R_α *(with* α *ranging over* Ags*) and* R_{Ags} *underlie* `Choice`. *Consider the following* unraveling variation $\mathcal{M}^u = \left\langle W^u, R^u_\Box, R^u_X, R^u_Y, \texttt{Choice}^u, \{\approx^u_\alpha\}_{\alpha \in Ags}, \mathcal{V}^u \right\rangle$*, defined as follows:*

- *Let* T^W *be the set of all finite sequences* $w_0, \dots, w_m$ *such that* $m \in \mathbb{N}$ *and* $w_i \in W$*. We define* $W^u \subseteq T^W \times \{0,1\}$*:*

  (a) $\langle w_0, \dots, w_m, 1 \rangle \in W^u$ *iff* $m \geq 0$ *and* $w_i R_X w_{i+1}$ *for all* $0 \leq i \leq m-1$.

  (b) $\langle w_0, \dots, w_m, 0 \rangle \in W^u$ *iff* $m > 0$ *and* $w_i R_Y w_{i+1}$ *for all* $0 \leq i \leq m-1$.

- *We define* $R^u_\Box$ *on* W^u *by the rule:* $\langle w_0, \dots, w_m, a \rangle \, R^u_\Box \, \langle v_0, \dots, v_l, b \rangle$ *iff* $a = b$*,* $l = m$*, and* $w_m R_\Box v_m$*.*

- *We define* R^u_X *on* W^u *by the rule:* $\langle w_0, \dots, w_m, a \rangle \, R^u_X \, \langle v_0, \dots, v_l, b \rangle$ *iff either*

  - $a = b = 1$*,* $m \geq 0$*,* $l = m+1$*,* $v_i = w_i$ *for every* $0 \leq i \leq m$*, and* $v_l = w_m^{+1}$*, or*
  - $a = b = 0$*,* $m > 1$*,* $l = m-1$*,* $v_i = w_i$ *for every* $0 \leq i \leq l$*, and* $v_l = w_m^{+1}$*, or*
  - $a = 0$*,* $b = 1$*,* $m = 1$*,* $l = 0$*,* $v_0 = w_0 = w_1^{+1}$*.*

- *We define* R^u_Y *on* W^u *by the rule:* $\langle w_0, \dots, w_m, a \rangle \, R^u_Y \, \langle v_0, \dots, v_l, b \rangle$ *iff either*

  - $a = b = 1$*,* $m > 0$*,* $l = m-1$*,* $v_i = w_i$ *for every* $0 \leq i \leq l$*, and* $v_l = w_m^{-1}$*, or*
  - $a = b = 0$*,* $m > 0$*,* $l = m+1$*,* $v_i = w_i$ *for every* $0 \leq i \leq m$*, and* $v_l = w_m^{-1}$*, or*
  - $a = 1$*,* $b = 0$*,* $m = 0$*,* $l = 1$*,* $w_0 = v_0 = v_1^{+1}$*.*

- *We define* $\texttt{Choice}^u$ *as follows:*

  - *For* $\alpha \in Ags$*, we define* R^u_α *on* W^u *by the rule:* $\langle w_0, \dots, w_m, a \rangle \, R^u_\alpha \, \langle v_0, \dots, v_l, b \rangle$ *iff* $a = b$*,* $l = m$*, and* $w_m R_\alpha v_m$*. We then set* $\texttt{Choice}^u_\alpha := \{R^u_\alpha[\langle w_0, \dots, w_m, a \rangle]; \langle w_0, \dots, w_m, a \rangle \in W^u\}$*, where* $R^u_\alpha[\langle w_0, \dots, w_m, a \rangle]$ *denotes the set of elements in the domain that are* R^u_α*-related to* $\langle w_0, \dots, w_m, a \rangle$*.*
  - *We define* R^u_{Ags} *on* W^u *by the rule:* $\langle w_0, \dots, w_m, a \rangle \, R^u_{Ags} \, \langle v_0, \dots, v_l, b \rangle$ *iff* $a = b$*,* $l = m$*, and* $w_m R_{Ags} v_m$*. We then set* $\texttt{Choice}^u_{Ags} = \left\{ R^u_{Ags}[\langle w_0, \dots, w_m, a \rangle]; \langle w_0, \dots, w_m, a \rangle \in W^u \right\}$*.*

- *For* $\alpha \in Ags$*, we define* $\approx^u_\alpha$ *on* W^u *by the rule:* $\langle w_0, \dots, w_m, a \rangle \approx^u_\alpha \langle v_0, \dots, v_l, b \rangle$ *iff* $w_m \approx_\alpha v_l$*.*

- *Recall that P is the set of propositions in $\mathcal{L}_{KX}$. We define $\mathcal{V}^m : P \to 2^{W^u}$ by the rule: $\langle w_0, \dots, w_m, a\rangle \in \mathcal{V}^u(p)$ iff $w_m \in \mathcal{V}(p)$.*

Proposition A.20. *If $\mathcal{M}$ is a super-additive Kripke-*exs*-n-model, then $\mathcal{M}^u$—as defined in Definition A.19—is a super-additive Kripke-*exs*-n-model where the transitive closures of R^u_X and R^u_Y are irreflexive.*

Proof. We want to show that $\left\langle W^u, R^u_\Box, R^u_X, R^u_Y, \mathtt{Choice}^u, \{\approx^u_\alpha\}_{\alpha\in Ags}\right\rangle$ is a super additive Kripke-*exs*-n-frame where the transitive closures of R^u_X and R^u_Y are irreflexive, which amounts to showing that the tuple satisfies the items in the definition of Kripke-*exs*-frames (Definition 3.9), that it satisfies the super-additivity, resp. cardinality-n, conditions, and that the transitive closures of R^u_X and R^u_Y are irreflexive.

1. It is routine to show that $R^u_\Box$ is an equivalence relation.

2. The fact that R_X, resp. R_Y, is serial and deterministic implies that R^u_X, resp. R^u_Y, is serial and deterministic. Observe that the variation of the traditional unraveling-argument, that we use here, plays an important role in ensuring that R^u_Y is serial: to have predecessors for one-element sequences, we introduced a construction that differentiates ascending from descending sequences. Therefore, for all $\langle w, 1\rangle \in W^u$, $\langle w, 1\rangle R^u_Y \left\langle w, w^{-1}, 0\right\rangle$. It is routine to show that $(\mathtt{Inverse})_{\mathrm{K}}$ holds, and the unraveling construction ensures that the transitive closures of R^u_X and R^u_Y are irreflexive.

3. We show that $\mathtt{Choice}^u$ fulfills the requirements of Definition 3.9 in their version for super-additive frames. It is routine to show that R^u_α and R^u_{Ags} are equivalence relations. Therefore, $\mathtt{Choice}^u$, as defined, indeed assigns a partition of W^u to each α, and to Ags. Super-additivity can be shown by proving that, for all $\alpha \in Ags$, $R^u_{Ags} \subseteq R^u_\alpha$. This follows from the definition of these relations and the fact that $R_{Ags} \subseteq R_\alpha$ for every $\alpha \in Ags$. Now, the definition of R^u_α implies that $(\mathtt{SET})_{\mathrm{K}}$ holds.

   $(\mathtt{IA})_{\mathrm{K}}$ Take $\langle w_0, \dots, w_m, a\rangle \in W^u$, and let $s : Ags \to 2^{\overline{\langle w_0, \dots, w_m, a\rangle}}$ be a function that maps each $\alpha \in Ags$ to a member of $\mathtt{Choice}^u_\alpha$ included in $\overline{\langle w_0, \dots, w_m, a\rangle}$. We want to show that $\bigcap_{\alpha\in Ags} s(\alpha) \neq \emptyset$. For $\alpha \in Ags$, take $\langle w_{\alpha 0}, \dots, w_{\alpha m}, a\rangle \in s(\alpha)$, then. We show that there exists $\langle v_0, \dots, v_m, a\rangle \in \overline{\langle w_0, \dots, w_m, a\rangle}$ such that $\langle v_0, \dots, v_m, a\rangle R^u_\alpha \langle w_{\alpha 0}, \dots, w_{\alpha m}, a\rangle$ for every $\alpha \in Ags$. Observe that the definition of s implies that $w_{\alpha m} \in \overline{w_m}$ for every $\alpha \in Ags$. Since $\mathcal{M}$ satisfies $(\mathtt{IA})_{\mathrm{K}}$, there exists $v_* \in \overline{w_m}$ such that $v_* R_\alpha w_{\alpha m}$ for every $\alpha \in Ags$. We have two cases, according to the value of a:

- (Case $a = 1$) The finite sequence given by $v_*^{-m}, \dots, v_*$ is such that $\langle v_*^{-m}, \dots, v_*, 1\rangle R_\alpha^u \langle w_{\alpha 0}, \dots, w_{\alpha m}, 1\rangle$ for every $\alpha \in Ags$.
- (Case $a = 0$) The finite sequence $v_*^{+m}, \dots, v_*$ is such that $\langle v_*^{+m}, \dots, v_*, 0\rangle R_\alpha^u \langle w_{\alpha 0}, \dots, w_{\alpha m}, 0\rangle$ for every $\alpha \in Ags$.

$(\mathtt{NAgs})_{\mathtt{K}}$ We want to show that $R_\Box^u \circ R_X^u \circ R_{Ags}^u \subseteq R_X^u \circ R_{Ags}^u$. Therefore, let $\langle w_0, \dots, w_m, a\rangle, \langle v_0, \dots, v_l, b\rangle \in W^u$ be such that $\langle w_0, \dots, w_m, a\rangle R_\Box^u \circ R_X^u \circ R_{Ags}^u \langle v_0, \dots, v_l, b\rangle$. We have three cases:

(i) (Case $a = b = 1, m \geq 0$) The assumption implies that there is $\left\langle w'_0, \dots, w'_m, 1\right\rangle \in W^u$ such that $\langle w_0, \dots, w_m, 1\rangle R_{Ags}^u \left\langle w'_0, \dots, w'_m, 1\right\rangle$ and $\left\langle w'_0, \dots, w'_{m+1}, 1\right\rangle R_\Box^u \langle v_0, \dots, v_l, 1\rangle$.[45] The first fact yields that $w_m R_{Ags} w'_m$, and the second implies by $(\mathtt{NAgs})_{\mathtt{K}}$ that $w'_m R_{Ags} v_m$. Therefore, transitivity of R_{Ags} yields that $w_m R_{Ags} v_m$. Then $\langle w_0, \dots, w_m, 1\rangle R_{Ags}^u \langle v_0, \dots, v_m, 1\rangle$, which gives us that $\langle w_0, \dots, w_m, 1\rangle R_X^u \circ R_{Ags}^u \langle v_0, \dots, v_{m+1}, 1\rangle$.

(ii) (Case $a = b = 0$, $m > 1$) Here, there is $\left\langle w'_0, \dots, w'_m, 0\right\rangle \in W^u$ such that $\langle w_0, \dots, w_m, 0\rangle R_{Ags}^u \left\langle w'_0, \dots, w'_m, 0\right\rangle$ and $\left\langle w'_0, \dots, w'_{m-1}, 0\right\rangle R_\Box^u \langle v_0, \dots, v_l, 0\rangle$. The first fact yields that $w_m R_{Ags} w'_m$, and the second implies by $(\mathtt{NAgs})_{\mathtt{K}}$ that $w'_m R_{Ags} v_l^{-1}$. Transitivity of R_{Ags} yields that $w_m R_{Ags} v_l^{-1}$. Then $\langle w_0, \dots, w_m, 0\rangle R_{Ags}^u \left\langle v_0, \dots, v_l, v_l^{-1}, 0\right\rangle$ and $\langle w_0, \dots, w_m, 0\rangle R_X^u \circ R_{Ags}^u \langle v_0, \dots, v_l, 0\rangle$.

(iii) (Case $a = 0$, $b = 1$, $m = 1$) The assumption implies that there exists $\left\langle w'_0, w'_1, 0\right\rangle \in W^u$ with $\langle w_0, w_1, 0\rangle R_{Ags}^u \left\langle w'_0, w'_1, 0\right\rangle$ and $\left\langle w'_0, 1\right\rangle R_\Box^u \langle v_0, 1\rangle$. The first fact yields that $w_1 R_{Ags} w'_1$, and the second implies by $(\mathtt{NAgs})_{\mathtt{K}}$ that $w'_1 R_{Ags} v_0^{-1}$. Therefore, transitivity of R_{Ags} yields that $w_1 R_{Ags} v_0^{-1}$, which implies that $\langle w_0, w_1, 0\rangle R_{Ags}^u \left\langle v_0, v_0^{-1}, 0\right\rangle$. In turn, this implies that $\langle w_0, w_1, 0\rangle R_X^u \circ R_{Ags}^u \langle v_0, 1\rangle$.

Finally, observe that, since for every $w \in W$ R_α and R_{Ags} induce partitions of cardinality at most n on $\overline{w}$, then R_α^u and R_{Ags}^u induce partitions of cardinality at most n on $\overline{\langle w_0, \dots, w_m, a\rangle}$ for every $\langle w_0, \dots, w_m, a\rangle$.

4. Observe that, for all $\alpha \in Ags$, $\approx_\alpha^u$, as defined, is an equivalence relation. We verify that $\mathcal{M}^u$ satisfies $(\mathtt{Unif-H})_{\mathtt{K}}$ and $(\mathtt{NoF})_{\mathtt{K}}$:

$(\mathtt{Unif-H})_{\mathtt{K}}$ Take $\alpha \in Ags$, and let $\langle v_0, \dots, v_m, a\rangle, \langle u_0, \dots, u_l, a'\rangle \in W^u$ be such that $\langle v_0, \dots, v_m, a\rangle \approx_\alpha^u \langle u_0, \dots, u_l, a'\rangle$, which implies that $(\star)$ $v_m \approx_\alpha u_l$. Take

[45] Observe that this last fact implies that $\langle v_0, \dots, v_l, 1\rangle = \langle v_0, \dots, v_{m+1}, 1\rangle$.

$\langle x_0,\dots,x_m,a\rangle \in \overline{\langle v_0,\dots,v_m,a\rangle}$. We want to show that there exists an element in $\overline{\langle u_0,\dots,u_l,a'\rangle}$ that is $\approx^u_\alpha$-related to $\langle x_0,\dots,x_m,a\rangle$. The fact that $\langle x_0,\dots,x_m,a\rangle \in \overline{\langle v_0,\dots,v_m,a\rangle}$ implies that $x_m \in \overline{v_m}$, so that ($\star$) and the fact that $\mathcal{M}$ satisfies $(\mathtt{Unif}-\mathtt{H})_{\mathtt{K}}$ entail that there exists $y \in \overline{u_l}$ such that $x_m \approx_\alpha y$. The finite sequence given by $y^{-l},\dots,y$ lies within T^W, so that $\left\langle y^{-l},\dots,y,a'\right\rangle \in \overline{\langle u_0,\dots,u_l,a'\rangle}$ and $\langle x_0,\dots,x_m,a\rangle \approx^u_\alpha \left\langle y^{-l},\dots,y,a'\right\rangle$.

$(\mathtt{NoF})_{\mathtt{K}}$ Take $\alpha \in Ags$, and let $\langle w_0,\dots,w_m,a\rangle, \langle v_0,\dots,v_l,b\rangle \in W^u$ be such that $\langle w_0,\dots,w_m,a\rangle \approx^u_\alpha \circ R^u_{\mathrm{X}} \langle v_0,\dots,v_l,b\rangle$. We want to show that $\langle w_0,\dots,w_m,a\rangle R^u_{\mathrm{X}} \circ \approx^u_\alpha \langle v_0,\dots,v_l,b\rangle$. We have three cases with three sub-cases each:

(Case $a=1, m\geq 0$) The assumption implies that $\langle w_0,\dots,w_{m+1},1\rangle \approx^u_\alpha \langle v_0,\dots,v_l,b\rangle$. We have the following sub-cases: ($b=1$, $l>0$) since $\mathcal{M}$ satisfies $(\mathtt{NoF})_{\mathtt{K}}$, $w_m \approx_\alpha v_{l-1}$, so that $\langle w_0,\dots,w_m,1\rangle \approx^u_\alpha \langle v_0,\dots,v_{l-1},1\rangle$; ($b=1$, $l=0$) here, $\mathcal{M}$'s $(\mathtt{NoF})_{\mathtt{K}}$ implies that $w_m \approx_\alpha v_0^{-1}$, so that $\langle w_0,\dots,w_m,1\rangle \approx^u_\alpha \left\langle v_0, v_0^{-1}, 0\right\rangle$; ($b=0$, $l>0$) here, $\mathcal{M}$'s $(\mathtt{NoF})_{\mathtt{K}}$ implies that $w_m \approx_\alpha v_{l+1}$, so that $\langle w_0,\dots,w_m,1\rangle \approx^u_\alpha \langle v_0,\dots,v_{l+1},0\rangle$.

(Case $a=0, m=1$) The assumption implies that $\langle w_0,1\rangle \approx^u_\alpha \langle v_0,\dots,v_l,b\rangle$. We have the following sub-cases: ($b=1$, $l>0$) here, $\mathcal{M}$'s $(\mathtt{NoF})_{\mathtt{K}}$ implies that $w_1 \approx_\alpha v_{l-1}$, so that $\langle w_0,w_1,0\rangle \approx^u_\alpha \langle v_0,\dots,v_{l-1},1\rangle$; ($b=1$, $l=0$) here, $\mathcal{M}$'s $(\mathtt{NoF})_{\mathtt{K}}$ implies that $w_1 \approx_\alpha v_0^{-1}$, so that $\langle w_0,w_1,0\rangle \approx^u_\alpha \left\langle v_0, v_0^{-1}, 0\right\rangle$; ($b=0$, $l>0$) here, $\mathcal{M}$'s $(\mathtt{NoF})_{\mathtt{K}}$ implies that $w_1 \approx_\alpha v_{l+1}$, so that $\langle w_0,w_1,0\rangle \approx^u_\alpha \langle v_0,\dots,v_{l+1},0\rangle$.

(Case $a=0, m>1$) The assumption implies that $\langle w_0,\dots,w_{m-1},0\rangle \approx^u_\alpha \langle v_0,\dots,v_l,b\rangle$. We have the following sub-cases: ($b=1, l>0$) here, $\mathcal{M}$'s $(\mathtt{NoF})_{\mathtt{K}}$ implies that $w_m \approx_\alpha v_{l-1}$, so that $\langle w_0,\dots,w_m,0\rangle \approx^u_\alpha \langle v_0,\dots,v_{l-1},1\rangle$; ($b=1$, $l=0$) here, $\mathcal{M}$'s $(\mathtt{NoF})_{\mathtt{K}}$ implies that $w_m \approx_\alpha v_0^{-1}$, so that $\langle w_0,\dots,w_m,0\rangle \approx^u_\alpha \left\langle v_0, v_0^{-1}, 0\right\rangle$; ($b=0$, $l>0$) here, $\mathcal{M}$'s $(\mathtt{NoF})_{\mathtt{K}}$ implies that $w_m \approx_\alpha v_{l+1}$, so that $\langle w_0,\dots,w_m,0\rangle \approx^u_\alpha \langle v_0,\dots,v_{l+1},0\rangle$.

□

Proposition A.21. *If $\mathcal{M}$ is a super-additive Kripke-*exs-*n-model, then $F : \mathcal{M}^u \to \mathcal{M}$, defined by $F(\langle w_0,\dots,w_m,a\rangle) = w_m$, is a surjective bounded morphism, where $\mathcal{M}^u$ is as defined in Definition A.19.*

Proof.

- By construction, F is surjective. Now, the definition of $\mathcal{V}^u$ in the last item of Definition A.19 ensures that $\langle w_0,\dots,w_m,a\rangle$ and $F(\langle w_0,\dots,w_m,a\rangle)$ satisfy the same propositional letters for every $\langle w_0,\dots,w_m,a\rangle \in W^u$.

- Take $R \in \left\{R_X, R_Y, R_\Box, R_\alpha, R_{Ags}, \approx_\alpha\right\}$, and let R^u stand for the corresponding relation on $\mathcal{M}^u$. Definition A.19 ensures that $\langle w_0, \ldots, w_m, a\rangle \, R^u \, \langle v_0, \ldots, v_l, b\rangle$ implies that $w_m R v_l$.

- Take $R \in \left\{R_X, R_Y, R_\Box, R_\alpha, R_{Ags}, \approx_\alpha\right\}$. Assume that $F(\langle w_0, \ldots, w_m, a\rangle) \, R \, v$ for some $v \in W$. We have the following cases:
  - (Case $a = 1$, $m > 0$) For $R \in \left\{R_\Box, R_\alpha, R_{Ags}, \approx_\alpha\right\}$, $\left\langle v^{-m}, \ldots, v^{-1}, v, 1\right\rangle$ is such that $\langle w_0, \ldots, w_m, 1\rangle \, R^u \left\langle v^{-m}, \ldots, v^{-1}, v, 1\right\rangle$. For R_X, $\langle w_0, \ldots, w_m, 1\rangle \, R^u_X \, \langle w_0, \ldots, w_m, v, 1\rangle$. For R_Y, $v = w_{m-1}$, so that $\langle w_0, \ldots, w_m, 1\rangle \, R^u_Y \, \langle w_0, \ldots, v, 1\rangle$.
  - (Case $a = 1$, $m = 0$) For $R \in \left\{R_\Box, R_\alpha, R_{Ags}, \approx_\alpha\right\}$, $\langle w_0, 1\rangle \, R^u \, \langle v, 1\rangle$. For R_X, $\langle w_0, 1\rangle \, R^u_X \, \langle w_0, v, 1\rangle$. For R_Y, $\langle w_0, 1\rangle \, R^u_Y \, \langle w_0, v, 0\rangle$.
  - (Case $a = 0$, $m = 1$) For $R \in \left\{R_\Box, R_\alpha, R_{Ags}, \approx_\alpha\right\}$, $\langle w_0, w_1, 0\rangle \, R^u \left\langle v^{+1}, v, 0\right\rangle$. For R_X, $w_0 = v$, and therefore $\langle w_0, w_1, 0\rangle \, R^u_X \, \langle v, 1\rangle$. For R_Y, $\langle w_0, w_1, 0\rangle \, R^u_Y \, \langle w_0, w_1, v, 0\rangle$.
  - (Case $a = 0$, $m > 1$) For $R \in \left\{R_\Box, R_\alpha, R_{Ags}, \approx_\alpha\right\}$, $\langle w_0, \ldots, w_m, 0\rangle \, R^u \left\langle v^{+m}, \ldots, v^{+1}, v, 0\right\rangle$. For R_X, $v = w_{m-1}$, so that $\langle w_0, \ldots, w_m, 0\rangle \, R^u_X \, \langle w_0, \ldots, v, 0\rangle$. For R_Y, $\langle w_0, \ldots, w_m, 0\rangle \, R^u_Y \, \langle w_0, \ldots, w_m, v, 0\rangle$.

Therefore, $F : \mathcal{M}^u \to \mathcal{M}$ is a surjective bounded morphism. □

Proposition A.22 (Completeness w.r.t. irreflexive super-additive n-models)**.** *The proof system Λ_{Kn} is complete with respect to the class of super-additive Kripke-*exs-*n*-models where the transitive closures of the 'next' and 'last' relations are irreflexive.*

Proof. Let φ be a Λ_{Kn}-consistent formula of $\mathcal{L}_{\mathsf{KX}}$. By Proposition A.18, there exists a super-additive Kripke-*exs-n*-model $\mathcal{M}$ and a world w in its domain such that $\mathcal{M}, w \models \varphi$. By Proposition A.21 and the invariance of modal satisfaction under bounded morphisms (Blackburn et al., 2002, Chapter 2), $\mathcal{M}^u$—as defined in Definition A.19—is a such that $\mathcal{M}^u, \langle w, 1\rangle \models \varphi$, where, by Proposition A.20, $\mathcal{M}^u$ is a super-additive Kripke-*exs-n*-model where the transitive closures of the 'next' and 'last' relations are irreflexive. □

Step 3: *Actual* Models

In the final step of our proof of completeness, for each Λ_{Kn}-consistent formula φ of $\mathcal{L}_{\mathsf{KX}}$, we build a model—in all the sense of the word—that satisfies it. As mentioned before, we adapt a well-known method from modal logic to produce a model where $R_{Ags} = \bigcap_{\alpha \in Ags} R_\alpha$ (Lorini, 2013; Schwarzentruber, 2012; Vakarelov, 1992). The idea is to build a structure—which I refer to as *matrix structure*—where the intersections of the agents' choice-cells include only one R_{Ags}-class. This is done by creating enough copies of worlds. The method would work just fine if our models only included the relations for historical necessity, individual and collective action, and the epistemic relations. However, the 'next' and 'last' relations, coupled with condition $(\texttt{NAgs})_{\texttt{K}}$, considerably complicate things. The process of copying worlds yields a big amount of them, so we need to be careful in drawing 'next' and 'last' relations between these copies. Step 2 comes in handy, because having a model in which the transitive closures of the 'next' and 'last' relations are irreflexive allows us to adequately define R_X and R_Y across the worlds' copies (see Footnote 46).

First, we introduce some auxiliary sets and terminology that will allow us to build the matrix structure.

Definition A.23. *Let* $\mathcal{M} = \left\langle W, R_\Box, R_X, R_Y, \texttt{Choice}, \{\approx_\alpha\}_{\alpha \in Ags}, \mathcal{V} \right\rangle$ *be a super-additive Kripke-*exs*-n-model where the transitive closures of the 'next' and 'last' relations are irreflexive and where equivalence relations R_α (with α ranging over Ags) and R_{Ags} underlie* `Choice`. *Consider the following definitions and remarks:*

- *Take $w \in W$ and $\alpha \in Ags$. Recall that the cardinality of partition $\texttt{Choice}^{\overline{w}}_{Ags}$ is bounded by n. Let A_w denote the set of all R_{Ags}-equivalence-classes included in $\bigcap\limits_{\alpha \in Ags} \texttt{Choice}^{\overline{w}}_\alpha(w)$. It is clear that, for all $v \in \overline{w}$, the cardinalities of the sets A_v are uniformly bounded by n, meaning that, for every $v \in \overline{w}$, card $(A_v) \leq n$. For $v \in \overline{w}$, let $\left\{c^v_0, \dots, c^v_{n-1}\right\}$ denote an enumeration of A_v with cardinality n (so that repetition is possible) such that, if $u \in \bigcap\limits_{\alpha \in Ags} \texttt{Choice}^{\overline{w}}_\alpha(v)$ (which happens iff $A_u = A_v$), then the enumeration $\left\{c^u_0, \dots, c^u_{n-1}\right\}$ of A_u is the same as $\left\{c^v_0, \dots, c^v_{n-1}\right\}$.*
- *For $w \in W$, $h[w] := \left\{v \in W; wR^*_Y v \text{ or } v = w \text{ or } wR^*_X v\right\}$, where R^*_Y, resp. R^*_X, is the transitive closure of R_Y, resp. R_X. Observe that, for all $w \in W$, R^*_X restricted to $h[w]$ is a strict linear order on $h[w]$, and that $h[v] = h[w]$ for every $v \in h[w]$. For $w \in W$, $h^-[w] := \left\{v \in h[w]; wR^*_Y v\right\}$, and $h^+[w] := \left\{v \in h[w]; wR^*_X v\right\}$. Observe that the fact that the transitive closure of R_X is irreflexive implies that, for all $w \in W$, $h^-[w]$, $\{w\}$, and $h^+[w]$ are pairwise disjoint sets.*

Lemma A.24. *Let $\mathcal{M}$ be a super-additive Kripke-*exs*-n-model where the transitive closures of the 'next' and 'last' relations are irreflexive, and where equivalence relations R_α (with α ranging over Ags) and R_{Ags} underlie* `Choice`. *Take $w, w' \in W$ such that $wR_\Box w'$. For each $v \in h^-[w]$, there exists $v' \in h^-[w']$ such that $vR_{Ags}v'$. And vice versa: for each $v' \in h^-[w']$, there exists $v \in h^-[w]$ such that $v'R_{Ags}v$.*

Proof. We show the result for w and assume an analogous argument for w'. We proceed by induction on j over the enumeration of $h^{-}[w]$ given by the set $\{w^{-1}, w^{-2}, \dots\}$. For w^{-1}, condition $(\texttt{NAgs})_{\texttt{K}}$ ensures that $w^{-1}R_{Ags}w'^{-1}$. Suppose that the property holds for w^{-j}, so that there exists $v' \in h^{-}[w']$ with $w^{-j}R_{Ags}v'$. Since $R_{Ags} \subseteq R_{\Box}$, this means that $w^{-j}R_{\Box}v'$. Since R_{Ags} is reflexive, this implies that $w^{-(j+1)}R_{\Box} \circ R_X \circ R_{Ags}v'$, which by $(\texttt{NAgs})_{\texttt{K}}$ implies that $w^{-(j+1)}R_{Ags}v'^{-1}$. □

Definition A.25 (Matrix structure). *Let* $\mathcal{M} = \left\langle W, R_{\Box}, R_X, R_Y, \texttt{Choice}, \{\approx_\alpha\}_{\alpha\in Ags}, \mathcal{V}\right\rangle$ *be a super-additive Kripke-*exs*-n-model where the transitive closures of the 'next' and 'last' relations are irreflexive, and where equivalence relations R_α (with α ranging over Ags) and R_{Ags} underlie* $\texttt{Choice}$*.*

We define $\mathcal{M}^m := \left\langle W^m, R^m_{\Box}, R^m_X, R^m_Y, \texttt{Choice}^m, \{\approx^m_\alpha\}_{\alpha\in Ags}, \mathcal{V}^m\right\rangle$*, a* matrix structure based on $\mathcal{M}$*, as follows:*

- *Let* $\prod\limits_{\alpha\in Ags}\{0,\dots,n-1\}$ *denote the cartesian product of set* $\{0,\dots,n-1\}$*, that consists of vectors of natural numbers between* 0 *and* $n-1$*. The domain* W^m *is given by*

$$W^m := \left\{ (\mathbf{f}, w)\,;\; \begin{array}{l} w \in W, \\ \mathbf{f} : h[w] \to \prod\limits_{\alpha\in Ags}\{0,\dots,n-1\} \textit{ maps worlds to vectors}, \\ \textit{for all } v \in h[w], v \in c^v_{\left(\sum_{\alpha\in Ags}(\mathbf{f}(v))_\alpha\right) \mod n} \\ \textit{for all } v, v' \in h^{-}[w] \textit{ s. t. } vR_{Ags}v', \mathbf{f}(v) = \mathbf{f}(v') \end{array} \right\}.$$

- *We define* $R^m_{\Box}$ *on* W^m *by the rule:* $(\mathbf{f}, w)\, R^m_{\Box}\, (\mathbf{f}', w')$ *iff* $wR_{\Box}w'$ *and, for all* $v \in h^{-}[w], v' \in h^{-}[w']$ *such that* $vR_{Ags}v'$*,* $\mathbf{f}(v) = \mathbf{f}'(v')$*.*

- *We define* R^m_X *on* W^m *by the rule:* $(\mathbf{f}, w)\, R^m_X\, (\mathbf{f}', w')$ *iff* wR_Xw' *and* $\mathbf{f}' = \mathbf{f}$*.*

- *We define* R^m_Y *on* W^m *by the rule:* $(\mathbf{f}, w)\, R^m_Y\, (\mathbf{f}', w')$ *iff* wR_Yw' *and* $\mathbf{f}' = \mathbf{f}$*.*

- *We define* $\texttt{Choice}^m$ *as follows:*

  - *For* $\alpha \in Ags$*, we define* R^m_α *on* W^m *by the rule:* $(\mathbf{f}, w)\, R^m_\alpha\, (\mathbf{f}', w')$ *iff*

    * $wR_\alpha w'$*,*
    * *for all* $v \in h^{-}[w], v' \in h^{-}[w']$ *such that* $vR_{Ags}v'$*,* $\mathbf{f}(v) = \mathbf{f}'(v')$*,*
    * $(\mathbf{f}'(w'))_\alpha = (\mathbf{f}(w))_\alpha$*, where* $(\cdot)_\alpha$ *denotes the denotes the* α^{th} *projection of the vector within the brackets.*

  We then set $\texttt{Choice}^m_\alpha = \{R^m_\alpha\,[(\mathbf{f}, w)]\,; (\mathbf{f}, w) \in W^m\}$*, where* $R^m_\alpha\,[(\mathbf{f}, w)]$ *denotes the set of elements in the domain that are* R^m_α*-related to* $(\mathbf{f}, w)$*.*

- $(\mathbf{f}, w)\, R^m_{Ags}\, (\mathbf{f}', w')$ *iff*
  - $*$ $wR_{Ags}w'$,
  - $*$ *for all* $v \in h^-[w], v' \in h^-[w']$ *such that* $vR_{Ags}v'$, $\mathbf{f}(v) = \mathbf{f}'(v')$,
  - $*$ $\mathbf{f}(w) = \mathbf{f}'(w')$.

  We then set $\texttt{Choice}^m_{Ags} = \left\{R^m_{Ags}\,[(\mathbf{f}, w)]\,; (\mathbf{f}, w) \in W^m\right\}$.

- *For* $\alpha \in Ags$, *we define* $\approx^m_\alpha$ *on* W^m *by the rule:* $(\mathbf{f}, w) \approx^m_\alpha (\mathbf{f}', w')$ *iff* $w \approx_\alpha w'$.

- *Recall that* P *is the set of propositions in* $\mathcal{L}_{\mathsf{KX}}$. *We define* $\mathcal{V}^m : P \to 2^{W^m}$ *by the rule:* $(\mathbf{f}, w) \in \mathcal{V}^m(p)$ *iff* $w \in \mathcal{V}(p)$.

We proceed to show that a matrix structure is an actual model.

Proposition A.26. *If* $\mathcal{M}$ *is a super-additive Kripke-*exs*-*n*-model where the transitive closures of the 'next' and 'last' relations are irreflexive, then* $\mathcal{M}^m$*—as defined in Definition A.25—is a Kripke-*exs*-*l*-model for some* $l \in \mathbb{N} - \{0\}$.

Proof. We want to show that $\left\langle W^m, R^m_\Box, R^m_X, R^m_Y, \texttt{Choice}^m, \{\approx^m_\alpha\}_{\alpha\in Ags}\right\rangle$ is a Kripke-*exs*-l-frame, which amounts to showing that the tuple satisfies the items in the definition of Kripke-*exs*-frames (Definition 3.9), that it satisfies cardinality-l conditions, and that the transitive closures of R^m_X and R^m_Y are irreflexive.

1. It is routine to show that $R^m_\Box$ is an equivalence relation. When checking for symmetry, resp. transitivity, with respect to the second condition in the definition of R^m_α—the one that demands that $(\mathbf{f}, w)\, R^m_\alpha\, (\mathbf{f}', w')$ only if, for all $v \in h^-[w], v' \in h^-[w']$ such that $vR_{Ags}v'$, $\mathbf{f}(v) = \mathbf{f}'(v')$—it follows from symmetry, resp. transitivity, of R_{Ags}.

2. Let us show that R^m_X and R^m_Y are serial and deterministic, and that their transitive closures are irreflexive. We show it for R^m_X and assume analogous arguments for R^m_Y. Take $(\mathbf{f}, w) \in W^m$. Since R_X is serial and deterministic, we know that there exists a unique $w^{+1} \in W$ such that wR_Xw^{+1}. Observe that $w^{+1} \in h[w]$, that $w^{+1} \in c^{w^{+1}}_{\left(\sum_{\alpha\in Ags}\left(\mathbf{f}\left(w^{+1}\right)\right)_\alpha\right) \bmod n}$, and that $h\left[w^{+1}\right] = h[w]$. This implies that the tuple $\left(\mathbf{f}, w^{+1}\right)$ is a member of W^m. Moreover, it is the *only* member of W^m such that $(\mathbf{f}, w)\, R^m_X \left(\mathbf{f}, w^{+1}\right)$. In turn, the assumption that the transitive closure of R_X is irreflexive straightforwardly implies that the transitive closure of R^m_X is irreflexive.

   For $(\texttt{Inverse})_{\mathsf{K}}$, take $(\mathbf{f}, w) \in W^m$, and let $\left(\mathbf{f}, w^{-1}\right)$, $\left(\mathbf{f}, w^{+1}\right)$ be the unique members of W^m such that $(\mathbf{f}, w)\, R^m_Y \left(\mathbf{f}, w^{-1}\right)$ and $(\mathbf{f}, w)\, R^m_X \left(\mathbf{f}, w^{+1}\right)$.

Assume that $\left(\mathbf{f}, w^{-1}\right) R^m_X (\mathbf{f}', w')$. Since $\mathcal{M}$ satisfies $(\texttt{Inverse})_{\mathrm{K}}$, $w' = w$. Now, by definition of R^m_X, the assumption that $\left(\mathbf{f}, w^{-1}\right) R^m_X (\mathbf{f}', w')$ implies that $\mathbf{f}' = \mathbf{f}$. Therefore, $(\mathbf{f}', w') = (\mathbf{f}, w)$, so that $R_X \circ R_Y = Id$. A similar argument yields that if $\left(\mathbf{f}, w^{+1}\right) R^m_Y (\mathbf{f}', w')$, then $(\mathbf{f}', w') = (\mathbf{f}, w)$, so that $R_Y \circ R_X = Id$.

3. We show that $\texttt{Choice}^m$ fulfills the requirements of Definition 3.9 and substantiates the fact that the frame underlying $\mathcal{M}$ is an actual frame.

It is routine to show that R^m_α is an equivalence relation for every $\alpha \in Ags$. Again, when checking for symmetry, resp. transitivity, with respect to the second condition in the definition of R^m_α—the one that demands that $(\mathbf{f}, w) R^m_\alpha (\mathbf{f}', w')$ only if, for all $v \in h^-[w], v' \in h^-[w']$ such that $vR_{Ags}v'$, $\mathbf{f}(v) = \mathbf{f}'(v')$—it follows from symmetry, resp. transitivity, of R_{Ags}. Therefore, $\texttt{Choice}^m_\alpha$ is indeed a partition of W^m for every $\alpha \in Ags$. Similarly, R^m_{Ags} is an equivalence relation. Therefore, $\texttt{Choice}^m_{Ags}$ is indeed a partition of W^m.

We now show that $\texttt{Choice}^m_{Ags}((\mathbf{f}, w)) = \bigcap\limits_{\alpha \in Ags} \texttt{Choice}^m_\alpha((\mathbf{f}, w))$ for every $(\mathbf{f}, w) \in W^m$. This amounts to showing that $R^m_{Ags} = \bigcap\limits_{\alpha \in Ags} R^m_\alpha$. For the $(\subseteq)$ inclusion, assume that $(\mathbf{f}, w) R^m_{Ags} (\mathbf{f}', w')$. This implies that $wR_{Ags}w'$, so the fact that $\mathcal{M}$ is super-additive yields that $wR_\alpha w'$ for every $\alpha \in Ags$. Our assumption also implies, by definition, that, for all $v \in h^-[w], v' \in h^-[w']$ such that $vR_{Ags}v'$, $\mathbf{f}(v) = \mathbf{f}'(v')$, and that $\mathbf{f}(w) = \mathbf{f}'(w')$. This implies that $(\mathbf{f}'(w'))_\alpha = (\mathbf{f}(w))_\alpha$ *for every* $\alpha \in Ags$. Therefore, $(\mathbf{f}, w) R^m_\alpha (\mathbf{f}', w')$ for every $\alpha \in Ags$, so that $R^m_{Ags} \subseteq \bigcap\limits_{\alpha \in Ags} R^m_\alpha$.

For the $(\supseteq)$ inclusion, assume that $(\mathbf{f}, w) R^m_\alpha (\mathbf{f}', w')$ for every $\alpha \in Ags$. This implies that $wR_\alpha w'$ for every $\alpha \in Ags$, that, for all $v \in h^-[w], v' \in h^-[w']$ such that $vR_{Ags}v'$, $\mathbf{f}(v) = \mathbf{f}'(v')$, and that $(\mathbf{f}'(w'))_\alpha = (\mathbf{f}(w))_\alpha$ for every $\alpha \in Ags$. Thus, $A_w = A_{w'}$ and $\mathbf{f}'(w') = \mathbf{f}(w)$, so that w and w' both lie within $c^w_{\left(\sum_{\alpha \in Ags}(\mathbf{f}(w))_\alpha\right) \bmod n}$ and thus $wR_{Ags}w'$. Therefore, $(\mathbf{f}, w) R^m_{Ags} (\mathbf{f}', w')$. With this we have shown that $R^m_{Ags} \supseteq \bigcap\limits_{\alpha \in Ags} R^m_\alpha$.

$(\texttt{SET})_{\mathrm{K}}$ Showing that $\mathcal{M}^m$ satisfies this condition is straightforward from the definitions of the R^m_α: $R^m_\alpha \subseteq R^m_\Box$ for every $\alpha \in Ags$.

$(\texttt{IA})_{\mathrm{K}}$ Take $(\mathbf{f}, w) \in W^m$, and let $s : Ags \to 2^{\overline{(\mathbf{f}, w)}}$ be a function that maps each $\alpha \in Ags$ to a member of $\texttt{Choice}^m_\alpha$ included in $\overline{(\mathbf{f}, w)}$. We want to show that $\bigcap_{\alpha \in Ags} s(\alpha) \neq \emptyset$. For each $s(\alpha)$, take $(\mathbf{f}^\alpha, w_\alpha) \in s(\alpha)$. For every $\alpha \in Ags$, the fact that $(\mathbf{f}^\alpha, w_\alpha) \in \overline{(\mathbf{f}, w)}$ implies that $w_\alpha \in \overline{w}$ and that $\mathbf{f}^\alpha(x_\alpha) = \mathbf{f}(x')$ for every $x_\alpha \in h^-[w_\alpha], x' \in h^-[w]$ such that $x_\alpha R_{Ags} x'$. Now, since $\mathcal{M}$ satisfies

$(\mathtt{IA})_{\mathtt{K}}$, $\bigcap_{\alpha \in Ags} \mathtt{Choice}^{\overline{w}}_{\alpha}(w_\alpha) \neq \emptyset$. Take $v'_* \in \bigcap_{\alpha \in Ags} \mathtt{Choice}^{\overline{w}}_{\alpha}(w_\alpha)$. Consider the number $N := \left(\sum_{\alpha \in Ags} (\mathbf{f}^{\alpha}(w_\alpha))_\alpha\right)$ mod n. Take $v_* \in c^{v'_*}_N$—where recall that $c^{v'_*}_N$ is an R_{Ags}-equivalence class in the enumeration of $A_{v'_*}$—the set of all R_{Ags}-equivalence classes included in $\bigcap_{\alpha \in Ags} \mathtt{Choice}^{\overline{v'_*}}_{\alpha}(v'_*) = \bigcap_{\alpha \in Ags} \mathtt{Choice}^{\overline{w}}_{\alpha}(w_\alpha)$. By construction, $v_* R_\Box w$ and $v_* R_\alpha w_\alpha$ for every $\alpha \in Ags$. We want to build a function $\mathbf{f}^*$ such that $(\mathbf{f}^*, v_*) \in \overline{(\mathbf{f}, w)}$ and such that $(\mathbf{f}^*, v_*) \in s(\alpha)$ for every $\alpha \in Ags$. In order to do this, we first observe that $\mathbf{f}^*$ should assign a vector to each $u \in h[v_*]$. We define $\mathbf{f}^*$ by parts and then show that our definition yields that $(\mathbf{f}^*, v_*)$ satisfies the conditions that we want. For all $u \in h^-[v_*]$, we know by Lemma A.24 that there exists $u' \in h^-[w]$ such that $u R_{Ags} u'$. Thus, for $u \in h^-[v_*]$, we set $\mathbf{f}^*(u) = \mathbf{f}(u')$. This part of f^* is well-defined because, for all $x, y \in h^-[w]$ such that $x R_{Ags} y$, $\mathbf{f}(x) = \mathbf{f}(y)$ (by definition of $\mathbf{f}$). For v_*, we set $\mathbf{f}^*(v_*) = \left((\mathbf{f}^{\alpha_1}(w_{\alpha_1}))_{\alpha_1}, \dots, (\mathbf{f}^{\alpha_m}(w_{\alpha_m}))_{\alpha_m}\right)$, where observe that $\left(\sum_{\alpha \in Ags} (\mathbf{f}^*(v))_\alpha\right)$ mod $n = \left(\sum_{\alpha \in Ags} (\mathbf{f}^{\alpha}(w_\alpha))_\alpha\right)$ mod $n = N$. Finally, for $u \in h^+[v_*]$, let N_u be the index of the element that includes u in the enumeration $\left\{c^u_0, \dots, c^u_{n-1}\right\}$ of A_u; if we take $(N_u, 0, \dots, 0) \in \prod_{\alpha \in Ags} \{0, \dots, n-1\}$, then $u \in c^u_{\left(\sum_{\alpha \in Ags} (N_u, 0, \dots, 0)_\alpha\right) \bmod n}$, where $(N_u, 0, \dots, 0)_\alpha$ denotes the α^{th} projection of vector $(N_u, 0, \dots, 0)$. Therefore, for $u \in h^+[v_*]$, we set $\mathbf{f}^*(u) = (N_u, 0, \dots, 0)$. Now, in order for f^* to be well-defined, *it is crucial* that the transitive closure of R_X is irreflexive, for this condition implies that $h^-[v_*], \{v_*\}$, and $h^+[v_*]$ are pairwise disjoint sets, as we had mentioned in Definition A.23.[46]

Observe that the definition of $\mathbf{f}^*$ implies that $(\mathbf{f}^*, v_*) \in W^m$. Let us show that it also implies that (a) $(\mathbf{f}^*, v_*) \in \overline{(\mathbf{f}, w)}$, and that (b) $(\mathbf{f}^*, v_*) \in s(\alpha)$ for every $\alpha \in Ags$. For (a), observe that we took $v_* \in \overline{w}$, and that we set $\mathbf{f}^*(u) = \mathbf{f}(u')$ for every $u \in h^-[v_*], u' \in h^-[w]$ such that $u R_{Ags} u'$, so that $(\mathbf{f}^*, v_*) \in \overline{(\mathbf{f}, w)}$. For (b), we have to prove the three items in the definition of R^m_α. Fix $\alpha \in Ags$. For the first item, observe that $v_* R_\alpha w_\alpha$. For the second item, we know that the fact that $(\mathbf{f}^\alpha, w_\alpha) \in \overline{(\mathbf{f}, w)}$ implies that, for $x_\alpha \in h^-[w_\alpha], x \in h^-[w]$ such that $x_\alpha R_{Ags} x$, $\mathbf{f}^\alpha(x_\alpha) = \mathbf{f}(x)$. Let $u \in h^-[v_*], u_\alpha \in h^-[w_\alpha]$ be such that $u R_{Ags} u_\alpha$. Lemma A.24

[46] Actually, this is the reason behind Step 2 in our proof of completeness: we use it to be able to *define* specific functions in the tuples that make up the domain of our structures, so that they are well-defined and so that the structure $\mathcal{M}^m$ satisfies the necessary conditions for it to be (a) an actual model and (b) a bounded morphic pre-image of the super-additive Kripke-*exs*-*n*-model that it is based on.

implies that $uR_{Ags}u'$ for some $u' \in h^{-}[w]$, so that point (a) above implies that $\mathbf{f}^{*}(u) = \mathbf{f}(u')$. The facts that $uR_{Ags}u'$ and $uR_{Ags}u_{\alpha}$ imply, by euclideanity of R_{Ags}, that $u'R_{Ags}u_{\alpha}$. Therefore, $\mathbf{f}^{\alpha}(u_{\alpha}) = \mathbf{f}(u')$. Thus, $\mathbf{f}^{*}(u) = \mathbf{f}(u') = \mathbf{f}^{\alpha}(u_{\alpha})$. Finally, for the third item, observe that $(\mathbf{f}^{*}(v_{*}))_{\alpha} = (\mathbf{f}^{\alpha}(w_{\alpha}))_{\alpha}$ by construction. Therefore, we have shown that $(\mathbf{f}^{*}, v_{*})\, R^{m}_{\alpha}\, (\mathbf{f}^{\alpha}, w_{\alpha})$ for every $\alpha \in Ags$, which means that $(\mathbf{f}^{*}, v_{*}) \in s(\alpha)$ for every $\alpha \in Ags$.

$(\mathtt{NAgs})_{\mathsf{K}}$ We want to show that $R^{m}_{\Box} \circ R^{m}_{X} \circ R^{m}_{Ags} \subseteq R^{m}_{X} \circ R^{m}_{Ags}$. Let $(\mathbf{f}, w), (\mathbf{g}, v) \in W^{m}$ be such that $(\mathbf{f}, w)\, R^{m}_{\Box} \circ R^{m}_{X} \circ R^{m}_{Ags}\, (\mathbf{g}, v)$, which means that there exists $(\mathbf{f}', w')$ such that $(\mathbf{f}, w)\, R^{m}_{Ags}\, (\mathbf{f}', w')$ and such that $\left(\mathbf{f}', w'^{+1}\right) R^{m}_{\Box}\, (\mathbf{g}, v)$. We want to show that $(\mathbf{f}, w)\, R^{m}_{Ags} \left(\mathbf{g}, v^{-1}\right)$. First, observe that the fact that $\left(\mathbf{f}', w'^{+1}\right) R^{m}_{\Box}\, (\mathbf{g}, v)$ implies that $w'^{+1} R_{\Box} v$, which by $(\mathtt{NAgs})_{\mathsf{K}}$ implies that $w' R_{Ags} v^{-1}$. It also implies that, for all $x \in h^{-}[w'^{+1}]$, $y \in h^{-}[v]$ such that $xR_{Ags}y$, $\mathbf{f}'(x) = \mathbf{g}(y)$. In particular, this means that $\mathbf{f}'(w') = \mathbf{g}\left(v^{-1}\right)$. Thus, our assumption that $(\mathbf{f}, w)\, R^{m}_{Ags}\, (\mathbf{f}', w')$ yields that two of the items in the definition of R^{m}_{Ags} (the first and third, respectively) hold between $(\mathbf{f}, w)$ and $\left(\mathbf{g}, v^{-1}\right)$, namely that $wR_{Ags}v^{-1}$ (via the facts that $wR_{Ags}w'$ and $w'R_{Ags}v^{-1}$) and that $\mathbf{f}(w) = \mathbf{f}'(w') = \mathbf{g}\left(v^{-1}\right)$. For the second item, we show that, for all $x \in h^{-}[w]$, $y \in h^{-}[v^{-1}]$ such that $xR_{Ags}y$, $\mathbf{f}(x) = \mathbf{g}(y)$: let $x \in h^{-}[w]$ and $y \in h^{-}\left[v^{-1}\right]$ be such that $xR_{Ags}y$; from the assumption that $(\mathbf{f}, w)\, R^{m}_{Ags}\, (\mathbf{f}', w')$ and Lemma A.24, there exists $x' \in h^{-}[w']$ such that $xR_{Ags}x'$ and $\mathbf{f}(x) = \mathbf{f}'(x')$, where $x' \in h^{-}[w']$ implies that $x' \in h^{-}\left[w'^{+1}\right]$ as well. Now, observe that the fact that $y \in h^{-}\left[v^{-1}\right]$ implies that $y \in h^{-}[v]$. In turn, by euclideanity of R_{Ags}, the facts that $xR_{Ags}y$ and that $xR_{Ags}x'$ imply that $yR_{Ags}x'$. Therefore, the elements $x' \in h^{-}\left[w'^{+1}\right]$ and $y \in h^{-}[v]$ are such that $yR_{Ags}x'$, so that the fact that $\left(\mathbf{f}', w'^{+1}\right) R^{m}_{\Box}\, (\mathbf{g}, v)$ implies that $\mathbf{f}'(x') = \mathbf{g}(y)$. In turn, this yields that $\mathbf{f}(x) = \mathbf{f}'(x') = \mathbf{g}(y)$, giving us what we wanted.

Finally, since for every $w \in W$ R_{α} and R_{Ags} both induce partitions of cardinality at most n on $\overline{w}$, then the construction of $\mathcal{M}^{m}$ ensures that R^{m}_{α} and R^{m}_{Ags} induce partitions of cardinality at most $n^{card(Ags)+1}$ on $\overline{(\mathbf{f}, w)}$ for every $(\mathbf{f}, w)$.

4. Since $\approx_{\alpha}$ is an equivalence relation on W for every $\alpha \in Ags$, $\approx^{m}_{\alpha}$ is also an equivalence relation on W^{m} for every $\alpha \in Ags$. We verify that $\mathcal{M}^{m}$ satisfies $(\mathtt{Unif-H})_{\mathsf{K}}$ and $(\mathtt{NoF})_{\mathsf{K}}$:

   $(\mathtt{Unif-H})_{\mathsf{K}}$ Let $(\mathbf{f}, w), (\mathbf{g}, v) \in W^{m}$ be such that $(\mathbf{f}, w) \approx^{m}_{\alpha} (\mathbf{g}, v)$. Take $(\mathbf{f}', w') \in \overline{(\mathbf{f}, w)}$. We want to show that there exists $(\mathbf{g}', v') \in \overline{(\mathbf{g}, v)}$ such that $(\mathbf{f}', w') \approx^{m}_{\alpha} (\mathbf{g}', v')$. By definition, the facts that $(\mathbf{f}', w') \in \overline{(\mathbf{f}, w)}$ and that $(\mathbf{f}, w) \approx^{m}_{\alpha} (\mathbf{g}, v)$

imply that $w'R_\Box w$ and that $w \approx_\alpha v$. Since $\mathcal{M}$ satisfies $(\mathtt{Unif-H})_\mathtt{K}$, there exists $v' \in \overline{v}$ such that $w' \approx_\alpha v'$. With a procedure similar to the one used to show that $(\mathtt{IA})_\mathtt{K}$ holds, we build a function $\mathbf{g}'$ such that $(\mathbf{g}', v') \in \overline{(\mathbf{g}, v)}$ and such that $(\mathbf{f}', w') \approx_\alpha^m (\mathbf{g}', v')$. We define $\mathbf{g}'$ by parts so that it satisfies the wanted conditions. For all $u' \in h^-[v']$, Lemma A.24 implies that there exists $u \in h^-[v]$ such that $u'R_{Ags}u$. Therefore, for $u' \in h^-[v']$, we set $\mathbf{g}'(u') = \mathbf{g}(u)$, where observe that this part of $\mathbf{g}'$ is well-defined because, for all $x, y \in h^-[v]$ such that $xR_{Ags}y$, $\mathbf{g}(x) = \mathbf{g}(y)$ (by definition of $\mathbf{g}$); for all $u' \in \{v'\} \cup h^+[v']$, there is some element in $A_{u'}$ that includes u'; let $N_{u'}$ be the index of such an element in the enumeration $\left\{c_0^{u'}, \dots, c_{n-1}^{u'}\right\}$; if we take $(N_{u'}, 0, \dots, 0) \in \prod\limits_{\alpha \in Ags} \{0, \dots, n-1\}$, then $u' \in c^{u'}_{\left(\sum_{\alpha \in Ags} (N_{u'}, 0, \dots, 0)_\alpha\right) \bmod n}$, where $(N_{u'}, 0, \dots, 0)_\alpha$ denotes the α^{th} projection of vector $(N_{u'}, 0, \dots, 0)$; therefore, for $u' \in \{v'\} \cup h^+[v']$, we set $\mathbf{g}'(u') = (N_{u'}, 0, \dots, 0)$. Again, for $\mathbf{g}'$ to be well-defined, it is crucial that the transitive closure of R_X is irreflexive, for this condition implies that $h^-[v']$ and $\{v'\} \cup h^+[v']$ are disjoint sets. Observe that the definition of $\mathbf{g}'$ implies that $(\mathbf{g}', v') \in W^m$, and that the fact that $w' \approx_\alpha v'$ implies, by definition of $\approx_\alpha^m$, that $(\mathbf{f}', w') \approx_\alpha^m (\mathbf{g}', v')$. By construction, $(\mathbf{g}', v') \in \overline{(\mathbf{g}, v)}$, since $v'R_\Box v$ and $\mathbf{g}'(u') = \mathbf{g}(u)$ for all $u' \in h^-[v'], u \in h^-[v]$ such that $u'R_{Ags}u$.

$(\mathtt{NoF})_\mathtt{K}$ Take $\alpha \in Ags$, and let $(\mathbf{f}, w), (\mathbf{g}, v) \in W^m$ be such that $(\mathbf{f}, w) \approx_\alpha^m \circ R_X^m (\mathbf{g}, v)$. We want to show that $(\mathbf{f}, w) R_X^m \circ \approx_\alpha^m (\mathbf{g}, v)$. By assumption, $\left(\mathbf{f}, w^{+1}\right) \approx_\alpha^m (\mathbf{g}, v)$. By definition of $\approx_\alpha^m$, this implies that $w^{+1} \approx_\alpha v$. Since $\mathcal{M}$ satisfies $(\mathtt{NoF})_\mathtt{K}$, then $w \approx_\alpha v^{-1}$. Therefore, the definition of $\approx_\alpha^m$ yields that $(\mathbf{f}, w) \approx_\alpha^m \left(\mathbf{g}, v^{-1}\right)$, and thus that $(\mathbf{f}, w) R_X^m \circ \approx_\alpha^m (\mathbf{g}, v)$.

□

Proposition A.27. *If $\mathcal{M}$ is a super-additive Kripke-*exs*-n-model where the transitive closures of the 'next' and 'last' relations are irreflexive, then $F : \mathcal{M}^m \to \mathcal{M}$, defined by $F((\mathbf{f}, w)) = w$, is a surjective bounded morphism, where $\mathcal{M}^m$ is as defined in Definition A.25.*

Proof.

- By construction, F is surjective. Now, the definition of $\mathcal{V}^m$ in the last item of Definition A.25 ensures that, for all $(\mathbf{f}, w) \in W^m$, $(\mathbf{f}, w)$ and $F((\mathbf{f}, w))$ satisfy the same propositional letters.

- Take $R \in \left\{R_X, R_Y, R_\Box, R_\alpha, R_{Ags}, \approx_\alpha\right\}$, and let R^m stand for the corresponding relation on $\mathcal{M}^m$. Definition A.25 ensures that the fact that $(\mathbf{f}, w) R^m (\mathbf{g}, v)$ implies that wRv.

- Assume that $F((\mathbf{f},w))\,R\,v$ for some $R \in \left\{R_X, R_Y, R_\Box, R_\alpha, R_{Ags}, \approx_\alpha\right\}$ and $v \in W$. We have the following cases:
  - (Case $R \in \{R_X, R_Y\}$) Observe that $(\mathbf{f}, v)$ is such that $(\mathbf{f},w)\,R^m\,(\mathbf{f},v)$.
  - (Case $R = R_\Box$). Assume that $F((\mathbf{f},w))\,R_\Box v$. This implies that $wR_\Box v$. Then $(\mathbf{g},v)$ is such that $(\mathbf{f},w)\,R^m_\Box\,(\mathbf{g},v)$, where we define $\mathbf{g}$ by parts as follows: for all $u \in h^-[v]$, we know by Lemma A.24 that there exists $u' \in h^-[w]$ such that $uR_{Ags}u'$; therefore, for $u \in h^-[v]$, we set $\mathbf{g}(u) = \mathbf{f}(u')$ (observe that this part of $\mathbf{g}$ is well-defined because, for all $x, y \in h^-[w]$ such that $xR_{Ags}y$, $\mathbf{f}(x) = \mathbf{f}(y)$, by definition of $\mathbf{f}$); for all $u \in \{v\} \cup h^+[v]$, there is some element in A_u that includes u; let N_u be the index of such an element in the enumeration $\left\{c^u_0, \dots, c^u_{n-1}\right\}$; if we take $(N_u, 0, \dots, 0) \in \prod_{\alpha \in Ags} \{0, \dots, n-1\}$, then $u \in c^u_{\left(\sum_{\alpha\in Ags}(N_u,0,\dots,0)_\alpha\right) \bmod n}$; therefore, for $u \in \{v\} \cup h^+[v]$, we set $\mathbf{g}(u) = (N_u, 0, \dots, 0)$. Once again, for $\mathbf{g}$ to be well-defined, it is crucial that the transitive closure of R_X is irreflexive, for this condition implies that $h^-[v]$ and $\{v\} \cup h^+[v]$ are disjoint sets. Observe that the definition of $\mathbf{g}$ implies that $(\mathbf{g},v) \in W^m$, so that the fact that $wR_\Box v$, coupled with our definition of $\mathbf{g}$, gives that $(\mathbf{f},w)\,R^m_\Box\,(\mathbf{g},v)$, since $\mathbf{g}(u) = \mathbf{f}(u')$ for all $u \in h^-[v], u' \in h^-[w]$ such that $uR_{Ags}u'$.
  - (Case $R = R_\alpha$). Fix $\alpha \in Ags$ and assume that $F((\mathbf{f},w))\,R_\alpha v$. This implies that $wR_\alpha v$. Observe, then, that $(\mathbf{g},v)$ is such that $(\mathbf{f},w)\,R^m_\alpha\,(\mathbf{g},v)$, where we define $\mathbf{g}$ by parts as follows: for $u \in h^-[v]$, we set $\mathbf{g}(u)$ exactly as we did in the above item for that part of the respective $\mathbf{g}$ that was defined over the preceding elements of the respective v; for v, we have to be careful; consider the elements in A_v; it is clear that there is an element in A_v that includes v; let N_v be the index of said element in the enumeration $\left\{c^v_0, \dots, c^v_{n-1}\right\}$, and let $M \in \mathbb{N}$ be such that $(M + (\mathbf{f}(w))_\alpha) \bmod n = N_v$; take $\alpha_1, \dots, \alpha_m$ an enumeration of Ags such that—without loss of generality—$\alpha = \alpha_j$, and take $(m_{\alpha_1}, \dots, m_{\alpha_m}) \in \prod_{\alpha\in Ags} \{0, \dots, n-1\}$ such that $m_{\alpha_j} = (\mathbf{f}(w))_\alpha$ and such that $\sum_{i\neq j} m_{\alpha_i} = M$; thus, $\left(\sum_{\alpha\in Ags}(m_{\alpha_1}, \dots, m_{\alpha_m})_\alpha\right) \bmod n = (M + (\mathbf{f}(w))_\alpha) \bmod n = N_v$; therefore, we set $\mathbf{g}(v) = (m_{\alpha_1}, \dots, m_{\alpha_m})$; for $u \in h^+[v]$, we define $\mathbf{g}(u)$ exactly as we did in the above item for that part of the respective $\mathbf{g}$ that was defined over the succeeding elements of the respective v. Once again, $\mathbf{g}$ is well-defined because the transitive closure of R_X is irreflexive, for this condition implies that $h^-[v]$, $\{v\}$, and $h^+[v]$ are pairwise disjoint sets. Observe that the definition of $\mathbf{g}$

implies that $(\mathbf{g}, v) \in W^m$. Moreover, the facts that (a) $wR_\alpha v$, that (b) $\mathbf{g}(u) = \mathbf{f}(u')$ for all $u \in h^-[v], u' \in h^-[w]$ such that $uR_{Ags}u'$, and that (c) $(\mathbf{g}(v))_\alpha = m_{\alpha_j} = (\mathbf{f}(w))_\alpha$ imply that $(\mathbf{f}, w)\, R^m_\alpha\, (\mathbf{g}, v)$.

- $\left(\text{Case } R = R_{Ags}\right)$ In this case, assume that $F((\mathbf{f}, w))\, R_{Ags}v$. This implies that $wR_{Ags}v$. Then $(\mathbf{g}, v)$ is such that $(\mathbf{f}, w)\, R^m_{Ags}\, (\mathbf{g}, v)$, where we define $\mathbf{g}$ by parts as follows: for $u \in h^-[v] \cup h^+[v]$, we set $\mathbf{g}(u)$ exactly as in the above item, for the parts of the respective $\mathbf{g}$ concerning the preceding and succeeding elements of the respective v. For v, we set $\mathbf{g}(v) = \mathbf{f}(w)$. Thus, $(\mathbf{f}, w)\, R^m_{Ags}\, (\mathbf{g}, v)$, by arguments analogous to the ones in the above items.
- (Case $R =\approx_\alpha$) Assume that $F((\mathbf{f}, w)) \approx_\alpha v$. Observe that $(\mathbf{f}, w) \approx^m_\alpha (\mathbf{g}, v)$ for any $\mathbf{g}$ that fulfills the requirements in Definition A.25 that would make $(\mathbf{g}, v)$ an actual element of W^m. One such $\mathbf{g}$ exists in virtue of an argument similar to the ones used above to build the respective $\mathbf{g}$'s over the succeeding elements of the respective v's.

Therefore, $F : \mathcal{M}^m \to \mathcal{M}$ is a surjective bounded morphism. □

Proposition A.28 (Completeness w.r.t actual models). *For all $n \in \mathbb{N} - \{0\}$, the proof system Λ_{Kn} is complete with respect to the class of Kripke-*exs*-n-models.*

Proof. Take $n \in \mathbb{N} - \{0\}$, and let φ be a Λ_{Kn}-consistent formula of $\mathcal{L}_{\mathsf{KX}}$. By Proposition A.22, there exists a super-additive Kripke-*exs*-*n*-model $\mathcal{M}$ where the transitive closures of the 'next' and 'last' relations are irreflexive, and a world w in its domain, such that $\mathcal{M}, w \models \varphi$. By Proposition A.27 and the invariance of modal satisfaction under bounded morphisms, the matrix structure $\mathcal{M}^m$—as defined in Definition A.25—is such that $\mathcal{M}^m, (\mathbf{f}, w) \models \varphi$, where Proposition A.26 yields that $\mathcal{M}^m$ is a Kripke-*exs*-*l*-model for some $l \in \mathbb{N} - \{0\}$. □

Therefore, the following result, appearing in the main body of the chapter, has been shown:

Theorem 3.11 (Soundness & Completeness of Λ_{Kn}). *For all $n \in \mathbb{N} - \{0\}$, the proof system Λ_{Kn} is sound and complete with respect to the class of Kripke-*exs*-n-models.*

4

Agency, Knowledge, and Obligation

' "Duty, conscience," they say—I'm not going to speak against duty and conscience, but how do we really understand them?'

Fyodor Dostoevsky, *Crime and Punishment*

Responsibility appears greater or lesser, depending on a greater or lesser knowledge of the conditions in which the man whose action is being reviewed found himself... and on the greater or lesser understanding of the causes of the act.

Leo Tolstoi, *War and Peace*

4.1 Introduction

Suppose that you are the leader of a rescuing team. Ten miners are trapped either in shaft A or in shaft B of a mine, but you do not know in which. An overflowing river threatens to flood the mine, and your rescuing team can deploy a huge sandbag so that it blocks one of the shafts, but not both. If one shaft is blocked, all the water will stream into the other shaft, killing any miners in it. If neither shaft is blocked, both will be partially flooded, and your calculations say that one miner will likely die in this case. As the leader of the rescuing team, you are asked what is to be done. What *should* you do?

The story above is part of a famous thought experiment known as the *Miners Paradox*. Put forward in an unpublished paper by Parfit (1988) but made popular by Kolodny and MacFarlane (2010), its analysis leads to opposing recommendations for the rescuers, according to the following reasoning: under a utilitarian view on morality, the goal would be to minimize the number of casualties, so that—considering the rescuers' uncertainty about the location of the miners—it seems correct to say that *(1) the rescuers ought to block neither shaft*; however, it also seems correct to say that *(2) if the miners are in shaft A, the rescuers ought to block shaft A,* and that *(3) if the miners are in shaft B, the rescuers ought to block shaft B*; obviously, the rescuers know that *(4) either the miners are in shaft A or they are in shaft B*; thus stated, sentences (2), (3), and (4) entail that *(5) the rescuers either ought to block shaft A or ought to block shaft B,* which contradicts sentence (1).

This conflict is based on the assumption that, from the aforementioned utilitarian perspective, sentence (1) is clearly true. However, the opposite of sentence (1) is concluded from the seemingly correct premises (2), (3), and (4). Thus, one derives a contradiction from apparently valid premises describing the rescuers' background knowledge.

The most popular approach for solving the problem has been through curbing the reasoning-by-cases principle that underlies the derivation with sentences (2), (3), and (4) (see, for instance, Willer, 2012). The challenge, then, is to come up with general criteria that would distinguish when the reasoning-by-cases should apply and when it should not. A different approach—which motivates the contents of this chapter—focuses on the rescuers' knowledge, targeting its implications on what they ought to do. This approach argues that clauses (2) and (3) are not entirely adequate and should be restated as follows: (2) *if the rescuers know that the miners are in shaft A, then they ought to block shaft A*, and (3) *if the rescuers know that the miners are in shaft B, then they ought to block shaft B.*

Rather than providing a solution to the *Miners Paradox*, the knowledge-based approach opens up a discussion about the interplay between agency, ought-to-do, and the knowledge that agents have when they choose some action. This chapter is devoted to studying such an interplay. The reason is that agency, knowledge, and ought-to-do's are three essential components of responsibility, according to the list introduced in Chapter 1 (p. 3). Recall from said list (see also Duijf, 2018, Chapters 2 & 3), that a common intuition in blame assignment is that an agent is blameworthy only if the agent failed to comply with one of its obligations. What happens, then, if the agent did not know how to comply? Should the agent be considered blameworthy? If it was impossible for the agent to have knowingly complied with an obligation, how much of an obligation was it in the first place? Based on the view that an appropriate formalization of the interplay between agency, knowledge, and ought-to-do can help us solve these questions, the main contribution of this chapter is the development of logics for two senses of ought-to-do: an *objective* sense and a *subjective* one.

Perhaps the best way to present these two senses is by means of an example. Consider once again the *Miners Paradox*. Its paradoxical character can be seen as a direct consequence of the assumption that the ought-to-do's appearing in sentences (1), (2), (3), and (5) are all of the same kind. I propose that, in fact, two senses of ought-to-do are being confused. Objectively speaking, sentence (1) is false! It is not the case that the rescuers should refrain from blocking either shaft, because they should block the shaft that contains all the miners and be done with it. Subjectively speaking, however, there is an argument in favor of saying that (5) is false: it is not the case that the rescuers should block one of the shafts, because they do not know which shaft needs to be blocked to save all the miners.[1]

I propose that an agent subjectively ought to do φ only when the agent has *practical knowledge* about the acts that would lead to φ (see the discussion on the concept of know-how in Chapter 3's Subsection 3.3.4). Since they do not know in which shaft the miners are trapped, the rescuers clearly lack the practical knowledge to save all ten miners. In other words, the rescuers cannot knowingly perform an action of the form 'to rescue all the miners by blocking a shaft.' The intuition, then, is that Kant's maxim of *ought implies can* (see Horty, 2001, Chapter 4) helps us draw a distinction between objective and subjective ought-to-do's: while objective ought-to-do's only require the possibility of successfully complying with them, subjective ones require the possibility of both successfully and knowingly complying with them.[2]

It is important to correctly position my approach—of differentiating two senses of ought-to-do—within the fast growing literature on issues involving agency, knowledge, and oughts (see, for instance, Baskent, Loohuis, & Parikh, 2012). I focus on what can be called *knowledge-dependent oughts*. These are somewhat similar to the *knowledge-based obligations* of Pacuit, Parikh, and Cogan (2006), but there is an important difference. While Pacuit et al. studied how acquiring knowledge about a state can lead agents to being obligated to do certain things, I study obligation from a perspective where agents should be excused for not complying with some duty if they lacked knowledge that is necessary for doing so.

[1] The latter view is certainly supported by the fact that the truth of sentence (1) is rarely called into question, something that must follow from a form of *subjective* reasoning. For instance, the truth of sentence (1) evokes the idea that the rescuers should make their choice on the basis of maximization of expected moral utility.

[2] Observe that the contrapositive of Kant's maxim can be stated as *cannot implies being excusable.* This type of excuse—not being able to comply with an obligation—is one of the most popular among a wide variety of excuses available for not doing something that one should do.

Thus, while Pacuit et al.'s approach concerns *theoretical knowledge* (if the doctors know that the patient is bleeding, then they ought to stop it), mine concerns *practical knowledge* (doctors ought to stop the bleeding of the patient, but if they do not know how to then they should be excused for not being able to stop it).

The starting point of my study comes from a recent interest in enhancing the expressivity (and the applicability) of Horty's (2001) seminal stit theory of ought-to-do so that it can deal with situations in which agents' knowledge plays a key role. Inspired by three puzzles for knowledge-dependent obligations, that pose problems for merely extending his *act-utilitarian stit theory* (*AUST*) with epistemic operators, Horty (2019) introduced a novel semantics for his so-called *epistemic oughts*.[3] In this chapter I carefully review Horty's (2019) proposal and build my own as a reply to it. Thus, my goal can be summarized as follows: I aim to provide axiomatizable stit-theoretic logics to reason about the reciprocity between three essential components of responsibility: agency, knowledge, and ought-to-do. Consequently, the logics that I develop (which are extensions of atemporal basic stit theory (*BST*) with both deontic and epistemic operators) have the following benefits: (a) their semantics for objective and subjective ought-to-do's deals with Horty's puzzles; (b) they pave the way for a simple formalization of responsibility; and (c) they are axiomatizable. An outline of this chapter is included below.

- Section 4.2 reviews the stit-theoretic background for the logics of objective and subjective ought-to-do's. Namely, the fundamentals of *AUST* and epistemic stit theory (*EST*) are addressed.
- Section 4.3 examines Horty's puzzles and the problems that they pose, pointing out the basic properties of Horty's own solution to these problems.
- Section 4.4 presents my stit logic of objective and subjective ought-to-do's, which I refer to as *epistemic act-utilitarian stit theory* (*EAUST*). The section shows how this logic admits a solution to the problems implied by Horty's puzzles, and it compares my proposal with Horty's aforementioned solution.
- Section 4.5 addresses *EAUST*'s logic-based and metalogic properties. Two Hilbert-style proof systems are introduced: one for the restriction that re-

[3] Indeed, these puzzles sum up both the conflict described for the *Miners Paradox* and the intuition that Kant's maxim of *ought implies can* should be tailored to epistemic cases. Section 4.3 thoroughly dissects these puzzles.

sults from considering only objective ought-to-do's, and the other for the restriction including only subjective ought-to-do's.[4] Importantly, this section presents soundness & completeness results for these proof systems.

- Section 4.6 (the conclusion) explores two paths for future work: first, it presents a doxastic sense of obligation, based on the extension of *EAUST* with a probabilistic semantics of belief; secondly, it discusses Horty's (2001, Chapter 6) group obligations and gives proposals both for their incorporation into *EAUST* and for their epistemic extension.

4.2 Agency, Knowledge, and Obligation in Stit Theory

As discussed in Chapter 2, stit theory was created to formalize agency. Thus, it naturally lent itself to the study both of *obligation* (Horty, 2001) and of *knowingly doing* (Broersen, 2011a). Since agency, knowledge, and obligation are three prominent components of responsibility, in this section I review the stit-theoretic literature on them to provide a bedrock for this chapter's discussions.[5]

Before this review, it is important to emphasize that—just as in all the other chapters of this thesis—the present description of the stit-theoretic modalities follows *my interpretation* of the semantics (see the discussion on p. 34 and Remark 2.4, p. 36). Therefore, when specifying the points of evaluation for the formulas—the indices in *bt*-models—I take it that at those indices states of affairs are definitive. Because of this, I use the present-perfect tense for the description of modality $[\alpha]\varphi$ and say that 'at index $\langle m, h\rangle$ α has seen to it that φ.' To be consistent, I use the past tense for modalities $\Box\varphi$ and $K_\alpha\varphi$ and say that 'at index $\langle m, h\rangle$ φ was settled' and that 'at index $\langle m, h\rangle$ α knew φ.' For modalities involving the verb 'ought,' whose only tense in English is the present tense (see Jörgensen, 1984), I use the past form of the sentences and say that 'at index $\langle m, h\rangle$ α ought to have seen to it that φ.' As discussed in Chapter 2, this usage does not mean to refer to past moments. Rather, it aims to reinforce the notion that, at the level of indices, circumstances in the world have already happened and cannot be changed.

[4] The reason for having these restrictions is two-fold: first, a proof system for the stit logic of objective ought-to-do's—which I presently call *AUST*—has already been given in the literature (Murakami, 2004). Secondly, the models with respect to which the proof systems that I present are sound and complete are precisely the models used in all the examples in this chapter. While I also have a soundness & completeness result for a merged proof system, the class of models with respect to which it is complete is much broader, and somewhat atypical. To achieve a proof of completeness for the full system, with respect to the models here used for the examples, is still an open problem.

[5] For a comprehensive review of these three concepts in the stit-theoretic literature, the reader is referred to Xu (2015).

The two backgrounds for the logics developed here are *act-utilitarian stit theory* (*AUST*) (see Chapter 2's Subsection 2.4.3) and *epistemic stit theory* (*EST*) (see Chapter 2's Subsection 2.4.4), where both these frameworks are extensions of *atemporal* BST (see Chapter 2's Section 2.3).

4.2.1 Act-Utilitarian Stit Theory

As for *AUST*, recall that it is the extension of atemporal *BST* that uses decision- and game-theoretic notions to formalize deontic concepts. This implies that *AUST* can be thought of as a *deontic logic.* Thus, let me briefly address the fundamentals of the deontic logic that serves as the broadest background for the ones that will be presented later on in this chapter.[6]

Standard deontic logic (*SDL*) is a modal logic with operators O for obligation and P for permission (Anderson, 1956; Kanger, 1970; Prior & Prior, 1955; von Wright, 1951). Modality $O\varphi$ is meant to express that it ought to be the case that φ, and $P\varphi$ is meant to express that φ is permitted. Indeed, P is the dual modal operator of O, such that P abbreviates $\neg O \neg$. The semantics for these modalities are given on standard Kripke models of the form $\mathcal{M} := \langle W, R_O, \mathcal{V} \rangle$, where W is a non-empty set of possible worlds, R_O is a serial relation dividing W into ideal and non-ideal worlds, and $\mathcal{V}$ is a valuation function assigning a set of worlds to each atomic proposition. Intuitively, the division into ideal and non-ideal worlds allows us to define what ought to be the case at some w as those formulas that hold in all the ideal alternatives of w. Thus, the evaluation rule for $O\varphi$ is as follows: $\mathcal{M}, w \models O\varphi$ iff for all w' such that wR_Ow', $\mathcal{M}, w' \models \varphi$.[7]

[6] Deontic logic is a broad field in symbolic modal logic. Starting with the seminal work by von Wright (1951), it has developed in many directions over the past 70 years. Horty (2001, Chapter 4) explicitly mentioned that, in the past, the task of relating deontic logic and act utilitarianism resulted in surprising difficulties, leading some authors (Castañeda, 1968, for instance) to suggest the possibility of a conflict in the fundamental principles underlying the two theories. According to Horty, one source of these difficulties is that, while deontic logic was developed as a theory of what ought or ought not *to be*, utilitarianism is concerned with ranking actions rather than states of affairs. Horty affirmed that stit theory can close this gap, because it allows us to develop a deontic logic to represent what agents ought *to do* under a particular variant of act utilitarianism, namely the *dominance theory* that will be presented in this chapter.

[7] This definition implies that O is a **KD** modal operator. By ordinary techniques of modal logic (see, for instance, Blackburn et al., 2002, Chapter 4), the logic of modality $O\varphi$ is then axiomatized by the proof system **KD**. For a discussion of *SDL* from a historical perspective, the reader is referred to Føllesdal and Hilpinen (1970). A more formal treatment, positioning *SDL* as a modal logic, was given by Chellas (1980, Chapter 6). As it happens, the literature has generally agreed on the fact that *SDL* suffers a number of paradoxes that hardly make it appropriate for deontic reasoning. As explained by Canavotto (2020, Chapter 6), one of the most serious issues involves the so-called *contrary-to-duty obligations*, which concern what ought to be done in the case where another, primary obligation has

Following Thomason (1981), Horty (2001, Chapter 3) adapted a version of *SDL* to branching-time logic and generalized it into a utilitarian setting. Horty's main goal was to provide a coherent backdrop for *AUST*, as well as to motivate his introduction of a formal, primitive notion of ought-to-do. In his adaptation, a function **Value** is added to a *bt*-model $\mathcal{M}$, such that **Value** assigns to each history a real number representing the deontic utility of the history. The idea, then, is that histories' utilities underlie which indices are better than others in $\mathcal{M}$, so that the formulas that hold at all 'good enough' indices are what ought to be the case. To clarify, the semantics on *bt*-models for $O\varphi$ is obtained by extending the recursive definition in Definition 2.3 with the following clause:

$$\mathcal{M}, \langle m, h\rangle \models O\varphi \quad \text{iff} \quad \begin{array}{l} \text{there is } h' \in H_m \text{ s. t. } \mathcal{M}, \langle m, h''\rangle \models \varphi \\ \text{for every } h'' \in H_m \text{ s. t. } \mathbf{Value}(h') \leq \mathbf{Value}(h''). \end{array} {}^{8}$$

Ever since deontic logic's inception there has been a debate about the relation between what ought to be the case and what agents ought to do. Actually, von Wright's (1951) seminal introduction of deontic operators treated these operators as applying not to propositions but rather to expressions representing actions. According to Horty (2001, Chapter 1), a variety of reasons—some of them purely technical—led other pioneers in the field (Anderson, 1956; Kanger, 1970; Prior & Prior, 1955, for instance) to adopt the more usual style of modal logic and apply the deontic operators to propositions. Thus, taking ought-to-be as primitive gained popularity in the literature, and with it the idea that the analysis of what agents ought to do is subsumed by the analysis of what ought to be the case, namely by identifying what an agent ought to do with what ought to be the case that the agent does.

However, it was also argued (see, for instance, Geach, 1991) that using logics of ought-to-be to reason about what agents ought to do might lead to "severe distortions" (Horty, 2001, p. 4). Indeed—albeit not for the same reasons that Geach advanced—this is exactly what happens if, with the semantics for $O\varphi$ on *bt*-models discussed above, one tries to characterize what agent α ought to do with the combined modality $O[\alpha]\varphi$ (thus identifying what agents ought to do with what ought to be the case that agents do). Referring to such an identification as the *Meinong/Chisholm analysis* of ought-to-do (because of the thesis expounded by

been violated. There are a number of paradoxes that revolve around contrary-to-duty obligations (see, for instance, Hilpinen & McNamara, 2013). Famously, Chisholm's (1963) paradox proved that *SDL* cannot properly handle contrary-to-duty obligations.

[8] With such a semantics, O turns out to be a **KD45** operator. Furthermore, formulas (a) $\Box\varphi \rightarrow O\varphi$, (b) $O\varphi \rightarrow \Box O\varphi$, and (c) $\neg O\varphi \rightarrow \Box\neg O\varphi$ are also valid with respect to the class of *bt*-models to which **Value** functions have been added.

Chisholm (1964)), Horty (2001, Chapter 3) explored its undesirable consequences in stit theory. To be precise, Horty proved that identifying ought-to-do with $O[\alpha]\varphi$ is vulnerable to *the gambling problem* (Example 4.1 below), which ultimately renders the Meinong/Chisholm analysis as inadequate to formalize ought-to-do.

Example 4.1 (The gambling problem)**.** The gambling problem *can be stated as follows: suppose that agent* Nikolai *is faced with two options at moment m_1: to either gamble 5 roubles or to forfeit the bet. If* Nikolai *gambles, then there is a history in which he wins 10 roubles, and another in which he loses his stake. However,* Nikolai *cannot determine whether he wins or loses. If* Nikolai *forfeits, then he preserves his original stake of 5 roubles no matter how things turn out. Thus, if the utility of each history at m_1 is entirely determined by the sum of money that* Nikolai *possessed in that history, the situation can be depicted by Figure 4.1.*

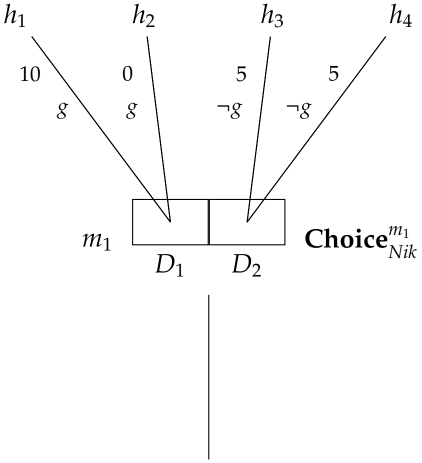


Figure 4.1: *The gambling problem.*

Here, D_1 represents the option of engaging in the gamble, and D_2 the option of forfeiting. The letter g, then, stands for the proposition 'Nikolai has gambled.' In this situation, a problem ensues due to the fact that, for all $h \in H_{m_1}$, $\mathcal{M}, \langle m_1, h\rangle \models O[Nik]g$: *at every index based on m_1 it ought to have been the case that* Nikolai *gambled.* Therefore, according to the Meinong/Chisholm analysis, *Nikolai* ought to have gambled, even if he risked achieving an outcome with payoff 0.

Horty (2001) developed his *AUST* as a way of solving this gambling problem (a robust objection against the Meinong/Chisholm analysis). Horty introduced a novel semantics for a primitive notion of ought-to-do, under von Wright's (1951) intuition that decoupling ought-to-do from ought-to-be is more appropriate for deontic reasoning. Instead of merely adapting *SDL* to branching-time logic, Horty used the theory of act utilitarianism, as well as decision-theoretic and game-theoretic ideas, to formalize his notion of ought-to-do.[9] The main definitions for his logic are included below.

Definition 4.2 (Syntax of *AUST*). *Given a finite set Ags of agent names and a countable set of propositions P, the grammar for the formal language $\mathcal{L}_O$ is given by*

$$\varphi ::= p \mid \neg\varphi \mid \varphi \wedge \varphi \mid \Box\varphi \mid [\alpha]\varphi \mid \odot_\alpha\varphi,$$

where p ranges over P and α ranges over Ags.

In this language, $\Box\varphi$ and $[\alpha]\varphi$ have the same meanings as in *BST* (see Definition 2.1, p. 28), and $\odot_\alpha\varphi$ is meant to express that agent α ought to have seen to it that φ. As for the semantics, the structures on which the formulas of $\mathcal{L}_O$ are evaluated are based on what I refer to as *act-utilitarian* bt-*frames.*

Definition 4.3 (*Aubt*-frames & models). *A tuple $\langle M, \sqsubset, Ags, \mathbf{Choice}, \mathbf{Value}\rangle$ is called an* act-utilitarian *bt*-frame *(*aubt-*frame for short) iff*

- $\langle M, \sqsubset, Ags, \mathbf{Choice}\rangle$ *is a* bt-*frame.*
- **Value** *is a function that assigns to each history $h \in H$ a real number, representing the deontic utility of h.*[10]

[9] In the words of Horty himself, "[t]he general goal of any utilitarian theory is to specify standards for classifying actions as right or wrong; and in its usual formulation, act utilitarianism defines an agent's action in some situation as right just in case the consequences of that action are at least as great in value as those of any of the alternatives open to the agent, and wrong otherwise" (Horty, 2001, p. 70). For a contemporary formulation of *act utilitarianism*, the reader is referred to Bergström (1966).

[10] While in game theory the utilities of outcomes quantify agents' *preferences*, in the context of *AUST* there is no particular interpretation for the word 'utility.' The utility of a history—**Value**(h)—does not stem from any agentive preference. Thus, *AUST* allows the assignment of values to "accommodate a variety of different approaches" (see Horty, 2001, Chapter 3, Section 2.2). To clarify, the notion of deontic utility is taken as primitive in *AUST*, and, rather than to individuals, it applies to the whole set of agents. Therefore, it may be thought of—but not necessarily so—as the "total utility of the set of agents in that history, their average utility, or perhaps some distribution-sensitive aggregation of the utilities of these individual agents" (Horty, 2001, p. 38). *AUST*'s analysis of interdependent decision contexts with *aubt*-frames thus differs from the game-theoretic approach in two main points: (1) rather than assigning individual utilities to each outcome as is done with the payoff functions in games, in *aubt*-frames function **Value** assigns a general deontic utility to each history; and (2) whereas in game theory utilities are assigned to full action profiles, in *aubt*-frames each index has a deontic utility. Of course, neither (1) nor (2) is strictly necessary for the formalization of ought-to-do. For instance, Kooi and Tamminga (2008) and Tamminga (2013) gave semantics for obligations on stit-theoretic models (known as *consequentialist models*) where individual utilities are assigned to each outcome and where a full action profile determines a single outcome (the latter being a property that I referred to as *determinism* in the context of the rich Kripke-*stit*-models of Definition 2.25, p. 67).

An aubt-*model* $\mathcal{M}$*, then, consists of the tuple that results from adding a valuation function* $\mathcal{V}$ *to an* aubt-*frame, where* $\mathcal{V} : P \rightarrow 2^{I(M \times H)}$ *assigns to each atomic proposition a set of indices.*

Horty (2001, Chapter 4) used *aubt*-models to provide semantics for $\odot_\alpha \varphi$. The intuition behind such a semantics is that the choices of action of a given agent can be ranked according to the utilities of the histories within them, and that ought-to-do's are based on the optimal choices in this ranking. To clarify, the idea is that at an index an agent ought to have seen to it that φ iff φ is an effect of all the agent's optimal choices. To formalize the ranking and the measure of optimality, Horty defined, for each moment and agent, a dominance ordering $\leq$ on the choices available to the agent at that moment. Of course, function **Value** is what underlies such a dominance ordering:

Definition 4.4 (Dominance ordering over choices)**.** *For an* aubt-*frame with* M *as its set of moments,* $\alpha \in Ags$*, and* $m \in M$*, consider the following definitions:*

- *Let* $\leq$ *be an ordering on* 2^{H_m} *defined by the rule: for* $X, Y \subseteq H_m$*,* $X \leq Y$ *iff* $\mathbf{Value}(h) \leq \mathbf{Value}(h')$ *for every* $h \in X$ *and* $h' \in Y$*. I write* $X < Y$ iff $X \leq Y$ *and* $Y \nleq X$*.*
- *Let* $\mathbf{State}^m_\alpha := \left\{S \subseteq H_m; S = \bigcap_{\beta \in Ags - \{\alpha\}} s(\beta), \text{ for } s \in \mathbf{Select}^m\right\}$*, where recall from Definition 2.2 (p. 29) that* $\mathbf{Select}^m$ *denotes the set of all selection functions at* m *(i.e., the functions that assign to each agent* β *a choice in* $\mathbf{Choice}^m_\beta$*).*
- *Let* $\leq$ *be an ordering on* $\mathbf{Choice}^m_\alpha$ *defined by the rule: for* $L, L' \in \mathbf{Choice}^m_\alpha$*,* $L \leq L'$ *iff for each* $S \in \mathbf{State}^m_\alpha$*,* $L \cap S \leq L' \cap S$*. I will refer to this ordering as the* dominance ordering over the choices available to α at *m, and, following Horty (2001), two types of dominance are defined using it:*
  - *L is* weakly dominated *by* L' *iff* $L \leq L'$*.*
  - *Let* $\prec$ *be defined by the rule: for* $L, L' \in \mathbf{Choice}^m_\alpha$*,* $L \prec L'$ *iff* $L \leq L'$ *and* $L' \nleq L$*. Then* L *is* strongly dominated *by* L' *iff* $L \prec L'$*.*
- *Let* $\mathbf{Optimal}^m_\alpha := \{L \in \mathbf{Choice}^m_\alpha; \text{there is no } L' \in \mathbf{Choice}^{m_*}_\alpha \text{such that } L \prec L'\}$. *This is the set of* α*'s optimal choices at* m*.*

An interesting question to ask is why one cannot use the ordering $\leq$ (the one in the first item in the definition above) to rank the choices. Why is the notion of $\mathbf{State}^m_\alpha$ introduced? Well, an important aspect of Horty's 'dominance act utilitarianism,' which is the name that he uses for the philosophical theory behind the definition above, is that the orderings are defined to account for *sure-thing reasoning*. A good way of discussing this kind of reasoning, and of therefore answering the questions regarding $\textbf{State}_{\alpha}^{m}$, is through an example:

Example 4.5 (Sure-thing reasoning). *Suppose that* Nikolai *and* Dolokhov *are playing a game at a gambling house. Both are holding a kopeck in one hand, and at moment m_1 both are independently faced with a choice between two actions: either placing their respective kopeck on a table heads up or placing it tails up. If both* Nikolai *and* Dolokhov *place their kopeck heads up, then the house gives them 10 roubles; if* Nikolai *places his kopeck heads up and* Dolokhov *places his tails up, the house pays 5 roubles; if* Nikolai *places his kopeck tails up and* Dolokhov *places his heads up, the house pays 9 roubles; finally, if both gamblers place their kopeck tails up, the house pays 4 roubles. Figure 4.2 includes a diagram for this situation.*

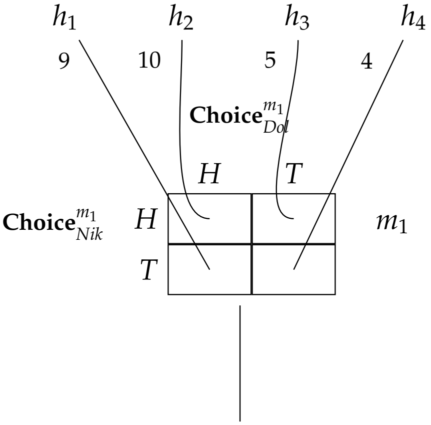


Figure 4.2: *Matching kopecks pt. 1: sure-thing reasoning.*

Here, the choices available to *Nikolai* at m_1 are placing his kopeck heads up (H) and placing it tails up (T). The choices available to *Dolokhov* at m_1 are the same as *Nikolai*'s. Thus, $\textbf{Choice}_{Nik}^{m_1} = \{H, T\}$, and $\textbf{Choice}_{Dol}^{m_1} = \{H, T\}$. Now, let us focus on *Nikolai*'s choices. Suppose for a second that, instead of using the dominance ordering $\leq$ of Definition 4.4 to rank these choices, one uses $\leq$. Observe, then, that $H \nleq T$ and that $T \nleq H$, so that if one were to use $\leq$ for the dominance ordering, neither H nor T would dominate the other choice. However, as Horty (2001, p. 63) wrote, "there seems to be a persuasive argument in favor of the conclusion that"

H is a better choice for *Nikolai* than *T*. To clarify, no matter whether *Dolokhov* performs *H* or *T*, it is better for *Nikolai* to perform *H* rather than *T*. This kind of argument is known as *sure-thing reasoning*.[11]

Therefore, as Example 4.5 suggests, using $\leq$ to rank the choices fails to capture sure-thing reasoning. Horty (2001, p. 63) wrote:

> The key to applying sure-thing reasoning in a given situation lies in identifying an appropriate partition of the possible outcomes into a set of states (sometimes called 'states of nature' or 'conditioning events'), against the background of which the actions available to an agent can then be evaluated through a state-by-state comparison of their results.

In *AUST*, such an appropriate partition of the outcomes is achieved by introducing the notion of $\mathbf{State}_{\alpha}^{m}$. This notion allows us to define the dominance ordering $\preceq$ so that sure-thing reasoning is accounted for. As the reader can notice, Definition 4.4 implies that, in Example 4.5, $T \prec H$ for *Nikolai*.[12]

With Definition 4.4's dominance ordering $\preceq$, then, the semantics for $\odot_{\alpha}\varphi$ is given as follows:

Definition 4.6 (Evaluation ruls for *AUST*). *Let $\mathcal{M}$ be an* aubt-*model. The semantics on $\mathcal{M}$ for the formulas of $\mathcal{L}_O$ are obtained by extending the recursive definition in Definition 2.3 (p. 30) with the following clause:*

$$\begin{array}{lll} \mathcal{M},\langle m,h\rangle \models \odot_{\alpha}\varphi & \textit{iff} & \text{for all } L \in \mathbf{Choice}_{\alpha}^{m} \text{ s. t. } \mathcal{M},\langle m,h_L\rangle \not\models \varphi \text{ for some } h_L \in L, \\ & & \text{there is } L' \in \mathbf{Choice}_{\alpha}^{m} \text{ s. t. } L \prec L' \text{ and, if } L'' = L \text{ or } L' \preceq L'', \\ & & \text{then } \mathcal{M},\langle m,h'\rangle \models \varphi \text{ for every } h' \in L''. \end{array}$$

[11] Sure-thing reasoning was first explicitly discussed by Savage (1954) (under the term 'sure-thing principle'), but it already appeared in an earlier work by the author (Savage, 1951). In the latter, he wrote that "there is one unquestionably appropriate criterion for preferring some act to some others: If for every possible state, the expected income of one act is never less and is in some cases greater than the corresponding income of another, then the former act is preferable to the latter" (Savage, 1951, p. 58).

[12] To account for sure-thing reasoning, the background states (against which choices are evaluated) must be independent of the available choices (see Horty, 2001, Chapter 4, pp. 64–67). Traditionally, there are two ways of thinking about this independence: (a) in terms of *probabilistic independence*, as explained by Jeffrey (1965), such that the conditional probability that a state holds should not vary according to the choice of action; and (b) in terms of *causal independence*, as first described by Gibbard and Harper (1978), such that the action chosen by an agent should not cause any effect that would influence the occurrence or absence of a state. It is based on the latter kind of independence—causal—that $\preceq$ is defined in *aubt*-models. The reason is that frame condition *independence of agency* (*IA*) implies that, for all $\alpha \in Ags$, $m \in M$, and state $S \in \mathbf{State}_{\alpha}^{m}$, $\mathbf{Choice}_{\alpha}^{m}(h) \cap S \neq \emptyset$ for every $h \in H_m$. In turn, this implies that S, seen as a set of alternative courses of events, can happen at m regardless of what α chooses. Thus, α's choices are always compatible with every state.

Satisfiability, validity, and general validity are defined as usual. I write $|\varphi|^m$ *to refer to the set* $\{h \in H_m; \mathcal{M}, \langle m,h\rangle \models \varphi\}$.

Therefore, one says that at index $\langle m,h\rangle$ agent α ought to have seen to it that φ iff for each action L of α that does not guarantee the bringing about of φ there is a better action L' such that (a) L' guarantees the bringing about of φ, and (b) every action that is better than L' also guarantees the bringing about of φ.

Before, I mentioned that Horty's intuitions involved a measure of optimality of choices. The trained reader will notice that the truth condition for $\odot_\alpha\varphi$ results from handling cases where the existence of an infinite number of available choices might yield that there are no optimal ones. Thus, such a truth condition can be much more intuitively stated when the definition of *aubt*-frames includes the requirement that the number of available choices at each moment is finite. In this case, Definition 4.6's clause for $\odot_\alpha\varphi$ is equivalent to: $\mathcal{M}, \langle m,h\rangle \models \odot_\alpha\varphi$ iff for all $L \in \mathbf{Optimal}_\alpha^m$, $\mathcal{M}, \langle m,h'\rangle \models \varphi$ for every $h' \in L$. It is in such terms that an agent's obligations are identified with the effects that are common to all the optimal choices for that agent.

4.2.2 Epistemic Stit Theory

As for *EST*, the reader is referred to Chapter 2's Subsection 2.4.4 for a discussion of its basic aspects. Furthermore, Chapter 3 was dedicated in its entirety to reviewing an extension of *EST* with temporal operators X and Y (for next and last moments, respectively), and all the ideas exposed there will greatly pay off throughout this chapter. To clarify, even if the present discussion is limited to agency with instantaneous effects, my treatment of knowledge will still involve the stages of information disclosure in the decision-making process, as well as the four kinds of knowledge formalized in Chapter 3. Therefore, with respect to the topics that are relevant in *EST* (see the list on p. 71), this chapter adopts the following conventions, applicable to any extension of an *ebt*-model $\mathcal{M}$ (see Definition 2.27, p. 70):

- Concerning *knowingly doing*, I will say that at index $\langle m,h\rangle$ agent α has knowingly seen to it that φ iff $\mathcal{M}, \langle m,h\rangle \models K_\alpha[\alpha]\varphi$—that is, iff at $\langle m,h\rangle$ α knew that it has seen to it that φ.

- Concerning the *epistemic sense of ability*, *know-how*, and *practical knowledge*, I will say that at $\langle m,h\rangle$ α was able in the epistemic sense to see to it that φ iff $\mathcal{M}, \langle m,h\rangle \models \Diamond K_\alpha[\alpha]\varphi$—that is, iff at $\langle m,h\rangle$ it was historically possible for α to knowingly see to it that φ.

- Concerning the *knowledge across the stages of information disclosure*, the exposition will be restricted to *ex ante* and *ex interim* knowledge. I will say that at $\langle m, h\rangle$ α had *ex ante* knowledge of φ iff $\mathcal{M}, \langle m, h\rangle \models \Box K_\alpha \varphi$—that is, iff at $\langle m, h\rangle$ it was historically settled that the agent knew φ. In turn, I will say that at $\langle m, h\rangle$ α had *ex interim* knowledge of φ iff $\mathcal{M}, \langle m, h\rangle \models K_\alpha[\alpha]\varphi$—that is, iff at $\langle m, h\rangle$ α has knowingly seen to it that φ.

- Concerning *uniformity*, all the models used here are built on the assumption that an agent should have the same available choices of action at indistinguishable indices. In the present context, the frame conditions characterizing this version of uniformity, as well as the formulas defining it, are thoroughly discussed in Sections 4.4 and 4.5.

This concludes the presentation of the two main backgrounds for this chapter's logics. With the notions covered, we are ready to start building an *appropriate* logic for studying the interplay between agency, knowledge, and ought-to-do. As implied before, the yardstick against which to measure the degree of appropriateness is given by two points: (a) the logics need to address the problems posed by Horty's (2019) puzzles; and (b) the logics need to be axiomatizable.

4.3 Horty's Puzzles

Horty (2019) presented three puzzles that are good examples of how the most natural mix between *AUST* and *EST* would approach the interplay between agency, knowledge, and ought-to-do. To scrutinize these puzzles in all formality, it is best to first address the question of what 'the most natural mix' actually is. Therefore, consider the following language, which results from extending $\mathcal{L}_O$ with epistemic modalities.

Definition 4.7 (Syntax for a possible extension). *Given a finite set Ags of agent names and a countable set of propositions P, the grammar for the formal language $\mathcal{L}_{\mathsf{KO}}$ is given by*

$$\varphi ::= p \mid \neg\varphi \mid \varphi \wedge \varphi \mid \Box\varphi \mid [\alpha]\varphi \mid K_\alpha\varphi \mid \odot_\alpha\varphi,$$

where p ranges over P and α ranges over Ags.

In this language, $\Box\varphi$ and $[\alpha]\varphi$ have the same meanings as in *BST* (see Definition 2.1, p. 28); $K_\alpha\varphi$ has the same meaning as in *EST* (Definition 2.26, p. 70); and $\odot_\alpha\varphi$ has the same meaning as in *AUST* (Definition 4.2). As for the semantics, for now let us think of the frames that result from combining *ebt*-frames

(Definition 2.27, p. 70) and *aubt*-frames (Definition 4.3) without any restriction, and allow me to refer to these structures as *unconstrained epistemic act-utilitarian branching-time frames*.[13]

Definition 4.8 (Unconstrained *eaubt*-frames)**.** *A tuple*

$$\left\langle M, \sqsubset, Ags, \mathbf{Choice}, \{\sim_\alpha\}_{\alpha \in Ags}, \mathbf{Value} \right\rangle$$

is called an unconstrained epistemic act-utilitarian branching-time frame *(unconstrained* eaubt-*frame for short) iff* $\left\langle M, \sqsubset, Ags, \mathbf{Choice}, \{\sim_\alpha\}_{\alpha \in Ags} \right\rangle$ *is an* ebt-*frame (Definition 2.27, p. 70) and* $\langle M, \sqsubset, Ags, \mathbf{Choice}, \mathbf{Value} \rangle$ *is an* aubt-*frame (Definition 4.3). An unconstrained* eaubt-*model* $\mathcal{M}$*, then, consists of the tuple that results from adding a valuation function* $\mathcal{V}$ *to an unconstrained* eaubt-*frame, where* $\mathcal{V} : P \to 2^{I(M \times H)}$ *assigns to each atomic proposition a set of indices.*

Definition 4.9 (Evaluation rules on unconstrained *eaubt*-models)**.** *Let* $\mathcal{M}$ *be an unconstrained* eaubt-*model. The semantics on* $\mathcal{M}$ *for the formulas of* $\mathcal{L}_{\mathsf{KO}}$ *are defined by extending the recursive definition in Definition 2.3 (p. 30) with the standard truth conditions for* $K_\alpha\varphi$ *and* $\odot_\alpha\varphi$ *(Definition 2.28 on p. 70 and Definition 4.6, respectively).*

Now, Horty's three puzzles posed problems for formalizing ought-to-do with the epistemic extension of *AUST* given by unconstrained *eaubt*-models. Let me address these puzzles one by one, then.

Example 4.10 (Puzzle #1)**.** *Suppose that* Nikolai *and* Dolokhov *are playing a game at a gambling house. The set-up is as follows:* Dolokhov *places a coin on top of a table—either heads up or tails up—and hides it from* Nikolai*.* Nikolai *can bet that the coin is heads up, bet that it is tails up, or forfeit the bet. If* Nikolai *bets and chooses correctly,* Nikolai *and* Dolokhov *win 10 roubles from the house. If he chooses incorrectly or forfeits the bet, they win nothing.*

To formalize Horty's interpretation of this puzzle, consider the unconstrained *eaubt*-model $\mathcal{M}$ depicted in Figure 4.3. Here, $Ags = \{Nikolai, Dolokhov\}$ and m_1, m_2, and m_3 are moments. At moment m_1 *Dolokhov* chooses between placing his coin on the table either heads up or tails up. Thus, his available actions are the following: D_1, where he places the coin heads up, and D_2, where he places the coin tails up. At moments m_2 and m_3 it is *Nikolai*'s turn to act, and his available actions are the following: N_1 and N_4, where he bets heads; N_2 and N_5, where he bets tails; and N_3 and N_6, where he forfeits the bet. The epistemic states of

[13] A specific sub-class of such unconstrained structures—resulting from two important frame conditions—will help us achieve an axiomatizable logic for the interplay of agency, knowledge, and ought-to-do. This sub-class is introduced in Section 4.4 (Definition 4.18).

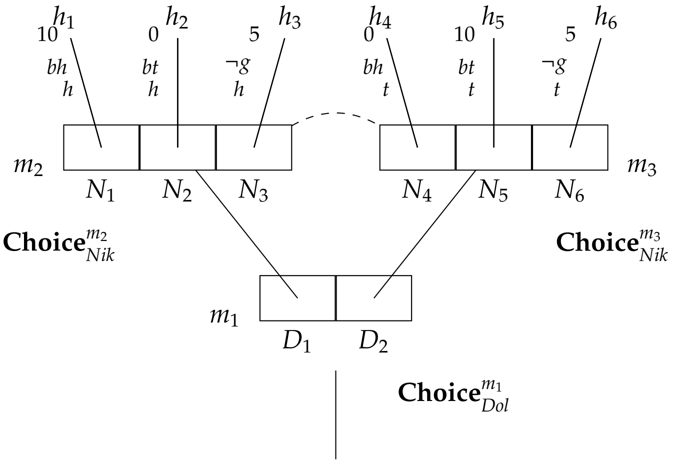


Figure 4.3: *Puzzle #1.*

Nikolai are represented with the indistinguishability relation given by the dashed line (where reflexive loops are omitted). Since *Dolokhov* is hiding his coin, *Nikolai* cannot distinguish whether he is at moment m_2 or at moment m_3. Thus, Horty set $\langle m_2, h\rangle \sim_{Nik} \langle m_3, h'\rangle$ for every $h \in H_{m_2}$ and $h' \in H_{m_3}$.

In Figure 4.3, h stands for the atomic proposition 'Dolokhov's coin is placed heads up,' t stands for 'Dolokhov's coin is placed tails up,' bh stands for 'Nikolai has bet heads,' bt stands for 'Nikolai has bet tails,' and g stands for 'Nikolai has gambled.' In such an interpretation, a problem ensues due to the fact that, for all $i \in \{2, 3\}$ and $h \in H_{m_i}$, $\mathcal{M}, \langle m_i, h\rangle \models K_{Nik} \odot_{Nik}\, g$: *at every index based on* m_2 *and* m_3, Nikolai *knew that he ought to have gambled*; however, to gamble is a risky move that could result in a payoff of 0, so that being obligated to gamble is counterintuitive, to say the least.[14]

Observe that the *Miners Paradox*—the example opening this chapter—can be modelled with a very similar diagram to Figure 4.3's, substituting *Nikolai* with agent *rescuer*, and setting the value of histories h_3 and h_6 at 9, the number of miners that *rescuer* saved at those histories. Let D_1 be the choice, available to the miners, of going into shaft A, and let D_2 be the choice of going into shaft B. Let N_1 and N_4 be the choices, available to *rescuer*, where she blocks shaft A of the mine; let N_2 and N_5 be the choices where she blocks shaft B; and let N_3 and N_6 be the choices where

[14] Observe that Horty's interpretation yields that *Nikolai* cannot knowingly perform any of his available actions. He cannot distinguish between the indices at which he has bet heads, at which he has bet tails, and at which he has forfeited the bet.

she refrains from blocking any shaft. Then the problem that was mentioned in the previous paragraph translates into a slightly terrifying conclusion: *at every index based on m_2 and m_3* rescuer *knew that she ought to have taken a gamble and blocked some shaft*. Therefore, the act-utilitarian stit-theoretic representation of the *Miners Paradox* is a variation of Example 4.10, and it is vulnerable to the problem that was raised for such an example.

Example 4.11 (Puzzle #2)**.** *Consider the same basic scheme as in Puzzle #1. If* Nikolai *bets and chooses correctly, he wins 10 roubles; however, this time, if he forfeits the bet, he* also *wins 10 roubles; if he bets incorrectly, he wins nothing. Horty's interpretation of this puzzle is depicted in Figure 4.4.*

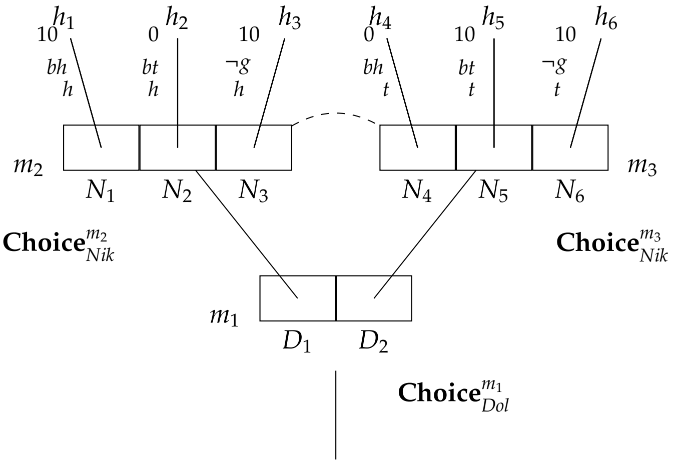


Figure 4.4: *Puzzle #2.*

Intuitively, *Nikolai* ought to have forfeited the bet. The reason is that by choosing to do so he would win by the same amount as when betting correctly, but without choosing an action that could possibly fail. The problem, then, is that, for all $i \in \{2,3\}$ and $h \in H_{m_i}$, $\mathcal{M}, \langle m_i, h\rangle \models \neg K_{Nik} \odot_{Nik} \neg g$: *at every index based on m_2 and m_3* Nikolai *did not know that he ought to have forfeited the bet.*

Example 4.12 (Puzzle #3)**.** *With the same basic scheme as in the previous puzzles, if* Nikolai *bets and chooses correctly, he wins 10 roubles. This time, if he bets incorrectly or forfeits the bet, he wins nothing. Horty's interpretation of this puzzle is depicted in Figure 4.5.*

In Figure 4.5, *w* stands for the proposition 'Nikolai and Dolokhov win.' Here, a problem ensues according to the following arguments: for all $i \in \{2,3\}$ and

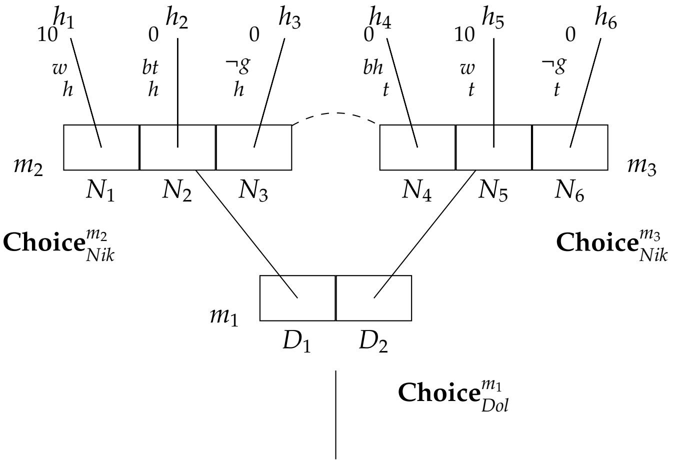


Figure 4.5: *Puzzle #3.*

$h \in H_{m_i}$, $\mathcal{M}, \langle m_i, h\rangle \models K_{Nik} \odot_{Nik} w$: *at all indices based on* m_2 *and* m_3 Nikolai *knew that he ought to have won;* however, having known that he ought to have won was not action-guiding in any sense, meaning that *Nikolai*'s knowledge of what he ought to have done could not influence any of his available choices. In terms of formulas, $\mathcal{M}, \langle m_i, h\rangle \not\models K_\alpha \odot_\alpha w \rightarrow \Diamond K_\alpha[\alpha]w$: *at all indices based on* m_2 *and* m_3 Nikolai *knew that he ought to have won, but it was impossible for him to knowingly win.* Thus, the epistemic version of Kant's principle of *ought implies can* is not satisfied.

4.3.1 Horty's Solution: Action Types

Introducing both syntactic and semantic addenda to epistemic *AUST*, Horty (2019) solved the problems implied by his three puzzles. On the syntactic side, he extended the language with modality $[\alpha\ \texttt{kstit}]\varphi$ (see Chapter 3's Subsection 3.3.2), meant to encode α's *ex interim* knowledge of φ or α's *epistemic agency* with respect to φ (Horty & Pacuit, 2017). On the semantic side, he based the semantics for $[\alpha\ \texttt{kstit}]\varphi$ on a set of action labels that was added to unconstrained *eaubt*-models. The premise behind this extension was that the only way to relate choices across indistinguishable indices was through tagging these choices with labels. To clarify, actions that have the same label may lead to different outcomes at different moments, and indistinguishable moments should offer choices with same labels (see Horty, 2019; Horty & Pacuit, 2017). Thus, in Horty's proposal such labels represent what is known in the literature as *action types*.[15]

To reason about types, one must also reason about *tokens*. As mentioned in Chapter 2's Subsection 2.4.2, the difference between types and tokens can be phrased as follows: while an action token is the single performance of an action by a specific agent at a specific moment, action types refer to categories or patterns of actions, that can be repeated at different moments by different agents, and that are instantiated in tokens.

A usual assumption in stit theory is that the cells in choice-partitions, that represent the actions available to an agent at some moment, are action tokens rather than types. The reason is that these cells are specific to each point in time and to each agent. Action types can then be seen as sets of tokens that share properties but may lead to different outcomes according to the moment when they—the tokens—are performed. Thus, extending the models with labels, meant to represent action types, accompanies the intent of binding several tokens together under a unifying term. Indeed, Horty's (2019) goal behind binding tokens together was to capture the uniformity condition that the same types of actions must be available at indistinguishable moments.[16] To be precise about his strategy, let me address the formal definitions for Horty's (2019) logic of epistemic oughts.

Definition 4.13 (Syntax for Horty's logic of epistemic oughts)**.** *Given a finite set Ags of agent names and a countable set of propositions P, the grammar for the formal language $\mathcal{L}_H$ is given by*

$$\begin{aligned} \varphi ::= \quad & p \mid \neg\varphi \mid \varphi \wedge \varphi \mid \Box\varphi \mid [\alpha\ \mathtt{stit}]\varphi \mid K_\alpha\varphi \mid [\alpha\ \mathtt{kstit}]\varphi \mid \\ & \odot[\alpha\ \mathtt{stit}]\varphi \mid \odot[\alpha\ \mathtt{kstit}]\varphi, \end{aligned}$$

where p ranges over P and α ranges over Ags.

In this language, $\Box\varphi$, and $K_\alpha\varphi$ have the same meanings as in the previous definitions of this chapter; $[\alpha\ \mathtt{stit}]\varphi$ is Horty's notation for $[\alpha]\varphi$; $[\alpha\ \mathtt{kstit}]\varphi$ is meant to express that agent α has seen to it that φ in an epistemic sense, or that

[15] Indeed, action types have a long-standing tradition in the literature on knowledge and agency, especially in the literature on *alternating-time temporal logic* (*ATL*) (Alur et al., 2002) and its epistemic extension, commonly referred to as *ATEL*. *ATEL*'s goal is to reason about the notion ability under imperfect information (see Ågotnes, 2006; Ågotnes et al., 2015; Jamroga & Ågotnes, 2007; van der Hoek & Wooldridge, 2002) (see also Subsections 2.4.2 and 2.4.4 of Chapter 2).

[16] In previous chapters I have referred to this condition as *uniformity of available action types* (UAAT) (see Footnote 40 on p. 65 and the discussion on p. 94, for instance).

α had *ex interim* knowledge of φ; $\odot[\alpha\ \texttt{stit}]\varphi$ is Horty's notation for $\odot_\alpha\varphi$; and $\odot[\alpha\ \texttt{kstit}]\varphi$ is meant to express that seeing to it that φ was α's epistemic ought. The formulas of this language are evaluated on what I refer to as *finite-choice Horty-like labelled* eaubt-*models*, whose definition is included below (Horty, 2019; Horty & Pacuit, 2017, notation adapted).

Definition 4.14 (Labelled *eaubt*-frames & models)**.** *A tuple of the form* $\left\langle M, \sqsubset, Ags, \textbf{Choice}, \{\sim_\alpha\}_{\alpha\in Ags}, \textbf{Value}, Tps, Lbl, Exe\right\rangle$ *is called a* labelled *eaubt*-frame *iff*

- $\left\langle M, \sqsubset, Ags, \textbf{Choice}, \{\sim_\alpha\}_{\alpha\in Ags}, \textbf{Value}\right\rangle$ *is an unconstrained* eaubt-*frame (Definition 4.8).*
- $\langle M, \sqsubset, Ags, \textbf{Choice}, Tps, Lbl, Exe\rangle$ *is a labelled* bt-*frame (Definition 2.24, p. 63).*

If in a labelled eaubt-*frame function* **Choice** *is such that* $\textbf{Choice}^m_\alpha$ *is finite for every* $\alpha \in Ags$ *and* $m \in M$*, then I will refer to the frame as a* finite-choice labelled *eaubt*-frame.[17] *If a labelled* eaubt-*frame additionally satisfies the following two constraints, then I will refer to the frame as a* Horty-like labelled *eaubt*-frame*:*

- (UAAT) *For every index* $\langle m, h\rangle$*, if* $\langle m, h\rangle \sim_\alpha \langle m', h'\rangle$*, then* $Tps^m_\alpha = Tps^{m'}_\alpha$*.*
- (C4) *For every index* $\langle m, h\rangle$*, if* $\langle m, h\rangle \sim_\alpha \langle m', h'\rangle$*, then* $\langle m, h_*\rangle \sim_\alpha \langle m', h'_*\rangle$ *for every* $h_* \in H_m$*,* $h'_* \in H_{m'}$*. This constraint allows us to lift the epistemic indistinguishability relations from the level of indices to the level of moments: in Horty-like labelled* eaubt-*frames, one writes* $m \sim_\alpha m'$ *iff there exist* $h \in H_m$ *and* $h' \in H_{m'}$ *such that* $\langle m, h\rangle \sim_\alpha \langle m', h'\rangle$*, and constraint* (C4) *implies that* $\sim_\alpha$ *is an equivalence relation on* M*.*[18]

A labelled *eaubt*-model $\mathcal{M}$*, then, is a tuple that results from adding a valuation function* $\mathcal{V}$ *to a labelled* eaubt-*frame, where* $\mathcal{V} : P \to 2^{I(M\times H)}$ *assigns to each atomic proposition a set of indices. If one adds a valuation like this to a tuple defining a finite-choice labelled frame, then I refer to the model as finite-choice. If one adds a valuation like this to a tuple defining a Horty-like labelled frame, then I refer to the model as Horty-like.*

Definition 4.15 (Dominance ordering of types)**.** *Let* $\mathcal{M}$ *be a finite-choice Horty-like labelled* eaubt-*model. For* $\alpha \in Ags$ *and* $m \in M$*, let* $\leq_H$ *be a dominance ordering on* Tps^m_α *defined by the following rule:* $\tau \leq_H \tau'$ *iff for all* m' *such that* $m \sim_\alpha m'$*,* $Exe^{m'}_\alpha(\tau) \leq$

[17] The terminology of 'finite-choice' also applies to any stit-theoretic frame for which the images of **Choice** are all finite partitions.

[18] The reader might find the tag for this constraint odd, but the reason is that I decided to refer to it with the same tag that Horty and Pacuit (2017) use.

$Exe^{m'}_{\alpha}(\tau')$*, where* $\leq$ *is the ordering defined in the first bullet point of Definition 4.4. I write* $\tau <_H \tau'$ iff $\tau \leq_H \tau'$ *and* $\tau' \not\leq_H \tau$*. The set of* α*'s optimal action types at* m *is then defined as* $TOptimal^{m}_{\alpha} := \{\tau \in Tps^{m}_{\alpha};$ there is no $\tau' \in Tps^{m}_{\alpha}$ s. t. $\tau <_H \tau'\}$.

Definition 4.16 (Evaluation rules for Horty's epistemic oughts)**.** *Let* $\mathcal{M}$ *be a finite-choice Horty-like labelled* eaubt-*model* $\mathcal{M}$*. The semantics on* $\mathcal{M}$ *for the formulas of* $\mathcal{L}_H$ *are defined recursively by the following truth conditions, evaluated at index* $\langle m,h\rangle$*:*

$$\begin{array}{lll}
\mathcal{M},\langle m,h\rangle \models p & \textit{iff} & \langle m,h\rangle \in \mathcal{V}(p) \\
\mathcal{M},\langle m,h\rangle \models \neg\varphi & \textit{iff} & \mathcal{M},\langle m,h\rangle \not\models \varphi \\
\mathcal{M},\langle m,h\rangle \models \varphi \wedge \psi & \textit{iff} & \mathcal{M},\langle m,h\rangle \models \varphi \textit{ and } \mathcal{M},\langle m,h\rangle \models \psi \\
\mathcal{M},\langle m,h\rangle \models \Box\varphi & \textit{iff} & \textit{for all } h' \in H_m, \mathcal{M},\langle m,h'\rangle \models \varphi \\
\mathcal{M},\langle m,h\rangle \models [\alpha\ \mathtt{stit}]\varphi & \textit{iff} & \textit{for all } h' \in \mathbf{Choice}^{m}_{\alpha}(h), \mathcal{M},\langle m,h'\rangle \models \varphi \\
\mathcal{M},\langle m,h\rangle \models K_{\alpha}\varphi & \textit{iff} & \textit{for all } \langle m',h'\rangle \textit{ s. t. } \langle m,h\rangle \sim_{\alpha} \langle m',h'\rangle, \\
 & & \mathcal{M},\langle m',h'\rangle \models \varphi \\
\mathcal{M},\langle m,h\rangle \models [\alpha\ \mathtt{kstit}]\varphi & \textit{iff} & \textit{for all } m' \textit{ s. t. } m \sim_{\alpha} m', \mathcal{M},\langle m',h'\rangle \models \varphi \\
 & & \textit{for every } h' \in H_{m'} \textit{ s. t. } Lbl_{\alpha}(\langle m,h\rangle) = Lbl_{\alpha}(\langle m',h'\rangle) \\
\mathcal{M},\langle m,h\rangle \models \odot[\alpha\ \mathtt{stit}]\varphi & \textit{iff} & \textit{for all } L \in Optimal^{m}_{\alpha}, \mathcal{M},\langle m,h'\rangle \models \varphi \textit{ for every } h' \in L \\
\mathcal{M},\langle m,h\rangle \models \odot[\alpha\ \mathtt{kstit}]\varphi & \textit{iff} & \textit{for all } \tau \in TOptimal^{m}_{\alpha}, \mathcal{M},\langle m',h'\rangle \models \varphi \textit{ for every} \\
 & & m' \textit{ s. t. } m \sim_{\alpha} m' \textit{ and every } h' \in Exe^{m'}_{\alpha}(\tau).
\end{array}$$

Thus, the main ideas underlying Horty's epistemic oughts are (a) that types can be ranked on the basis of an ordering of tokens at indistinguishable states, and (b) that epistemic oughts are effects of all the tokens that result from executing the optimal types at indices that are indistinguishable from the one of evaluation. In other words, at a given index to have seen to it that φ was an epistemic ought of agent α iff φ is an effect of all tokens that result from the execution of the types in $TOptimal^{m}_{\alpha}$.

As the reader can quickly verify (see Horty, 2019, for details), these definitions solve the problems posed by the puzzles introduced in the previous subsection. Let me briefly elaborate on such a solution.

In Examples 4.10, 4.11, and 4.12, *Nikolai*'s available actions—action tokens, that is—at m_2 are of the same type as those available at m_3. To clarify, consider once again Figures 4.3, 4.4, and 4.5. For *Nikolai*'s choices N_1 and N_4, where he bets heads, Horty set that $Lbl(N_1) = Lbl(N_4)$, so that N_1 and N_4 are two action tokens of the type 'to bet heads.' For choices N_2 and N_5, where *Nikolai* bets tails, $Lbl(N_2) = Lbl(N_5)$; and for choices N_3 and N_6, where *Nikolai* forfeits the bet, $Lbl(N_3) = Lbl(N_6)$.

Thus, interpreting Example 4.10 (Puzzle #1) as a Horty-like finite-choice labelled *eaubt*-model $\mathcal{M}$, the definition of $\preceq_H$ implies that $Lbl(L_i) \in TOptimal^{m_j}_{Nik}$ for every $3 \leq i \leq 8$ and $j \in \{2,3\}$. Thus, for all $j \in \{2,3\}$ and $h \in H_{m_j}$, $\mathcal{M}, \langle m_j, h \rangle \models \neg \odot [Nik \ \texttt{kstit}]g$: Nikolai *was not epistemically obligated to gamble*, which implies that $\mathcal{M}, \langle m_j, h \rangle \models K_{Nik} \neg \odot [Nik \ \texttt{kstit}]g$: Nikolai *knew that he was not epistemically obligated to gamble*. In Example 4.11 (Puzzle #2), $TOptimal^{m_j}_{Nik} = \{Lbl(N_3)\}$ for all $j \in \{2,3\}$. Thus, for all $j \in \{2,3\}$ and $h \in H_{m_j}$, $\mathcal{M}, \langle m_j, h \rangle \models K_{Nik} \odot [Nik \ \texttt{kstit}]\neg g$: Nikolai *knew that he was epistemically obligated to not gamble*. In Example 4.12 (Puzzle #3), $Lbl(L_i) \in TOptimal^{m_j}_{Nik}$ for every $3 \leq i \leq 8$ and $j \in \{2,3\}$. Thus, for all $j \in \{2,3\}$ and $h \in H_{m_j}$, $\mathcal{M}, \langle m_j, h \rangle \models \neg \odot [Nik \ \texttt{kstit}]w$: Nikolai *was not epistemically obligated to win*, which implies that $\mathcal{M}, \langle m_j, h \rangle \models K_{Nik} \neg \odot [Nik \ \texttt{kstit}]w$: Nikolai *knew that he was not epistemically obligated to win*.

As a solution to the puzzles' problems, then, Horty's approach is successful. However, one can raise two points of criticism:

1. Horty and Pacuit (2017) explained that constraint (C4) was imposed in the definition of labelled *eaubt*-models (Definition 4.14) just so that $[\alpha \ \texttt{kstit}]$ would result in an **S5** operator. Before, I mentioned that (C4) implies that the indistinguishability relations characterize uncertainty at the level of moments, rather than at the level of indices. The problem with (C4), then, is that it limits the class of models to those in which knowledge is moment-dependent. When knowledge is moment-dependent, agents cannot know whether they performed one non-trivial action instead of another. To clarify, formula $K_\alpha \varphi \rightarrow \Box \varphi$ is valid in Horty's (2019) logic, so that agents can only know things that are settled.[19]

   Similarly—and as mentioned in Chapter 3 (p. 94)—constraint (UAAT) and the semantics for $[\alpha \ \texttt{kstit}]\varphi$ entail that Horty's logic satisfies what I previously referred to as *knowledge of one's own action* (KOA) (see the discussion about this property on p. 92, as well as Footnote 19 on p. 93). According to (KOA), agents cannot have uncertainty about the actions that they perform at the *ex interim* stage. In other words, if p^τ_α denotes the proposition 'the action type τ is performed by agent α,' then formula $p^\tau_\alpha \rightarrow [\alpha \ \texttt{kstit}]p^\tau_\alpha$ is valid.

2. Horty and Pacuit (2017) argued that action types are necessary to deal with action tokens across indistinguishable states. As shown by Duijf et al. (2021) in the context of the stit-theoretic formalization of the epistemic sense of ability (and as will be shown in Section 4.4 in the context of the relation between knowledge and ought-to-do), this is not the case. To be precise, in Subsection 4.4.3 a general correspondence result is put forward,

[19] It is precisely because of this that Horty identified the kind of knowledge that $K_\alpha \varphi$ expresses in his logic as *ex ante* knowledge (see Chapter 3's Subsection 3.3.1).

> that proves that any labelled *eaubt*-model—which includes action types—can be transformed into a specific kind of *eaubt*-model—which eschews action types—such that (a) epistemic agency in the former corresponds to knowingly doing in the latter, and such that (b) epistemic oughts in the former correspond to subjective ought-to-do's in the latter. Thus, the introduction of action types leads to unnecessary complications both in the models and in the truth conditions for $[\alpha \texttt{ kstit}]\varphi$ and $\odot[\alpha \texttt{ kstit}]\varphi$.

Summing up, although Horty's approach is successful as a solution to the puzzles' problems, one can also be successful without using action types! Supporting this last claim is one of the main objectives of the next section.

4.4 A Logic of Objective & Subjective Oughts

In this important section I present a framework that solves the problems implied by the puzzles of Section 4.3 and that, when compared with Horty's (2019) solution, includes the following advantages: (a) it offers simpler semantics (without action types) that are more naturally related to Horty's (2001) seminal theory of ought-to-do (Definitions 4.18–4.21); (b) it is more flexible, since the study of knowledge, agency, and ought-to-do is not as limited by model constraints as in Horty's (2019) approach; and (c) it admits more straightforward syntactic characterizations, something that is very important for axiomatization (see Section 4.5).[20]

Following my joint works with Jan Broersen (Abarca & Broersen, 2019; Broersen & Abarca, 2018a, 2018b), I propose to disambiguate two senses of ought-to-do: an *objective* one, which coincides with Horty's act-utilitarian ought-to-do, and a *subjective* one, which arises from the epistemically best candidates in the set of available choices for an agent. By 'epistemically best' I mean those choices that are undominated not only at the index of evaluation but whose epistemic equivalents across indistinguishable indices are also undominated. Thus, we are talking about an extension of the language in Definition 4.7 with a new modality $\odot^S_\alpha\varphi$, meant to express that agent α subjectively ought to have done φ. The basic idea is that while $\odot_\alpha\varphi$ is correlated with α's causal agency and causal ability, $\odot^S_\alpha\varphi$ is correlated with α's knowingly-doing and ability in the epistemic sense. Let me cut to the chase and go ahead with the formal definitions.

[20] Whether Horty's (2019) logic of epistemic oughts is axiomatizable is still an open problem.

4.4.1 Syntax & Semantics

Definition 4.17 (Syntax for epistemic act-utilitarian stit theory (*EAUST*))**.** *Given a finite set Ags of agent names and a countable set of propositions P, the grammar for the formal language $\mathcal{L}_{\mathsf{KO}}$ is given by*

$$\varphi ::= p \mid \neg\varphi \mid \varphi \wedge \varphi \mid \Box\varphi \mid [\alpha]\varphi \mid K_\alpha\varphi \mid \odot_\alpha\varphi \mid \odot_\alpha^S\varphi,$$

where p ranges over P and α ranges over Ags.

In this language, $\Box\varphi$, $[\alpha]\varphi$, and $K_\alpha\varphi$ have the same meanings as in previous definitions; $\odot_\alpha\varphi$ is the same modality as in *AUST*, but I reinterpret it as expressing that agent α objectively ought to have seen to it that φ; and $\odot_\alpha^S\varphi$ is meant to express that α subjectively ought to have seen to it that φ. As for the semantics, the structures on which the formulas of $\mathcal{L}_{\mathsf{KO}}$ are evaluated are based on (constrained) *eaubt*-frames:

Definition 4.18 (*Eaubt*-frames & models)**.** *A tuple*

$$\left\langle M, \sqsubset, Ags, \mathbf{Choice}, \{\sim_\alpha\}_{\alpha\in Ags}, \mathbf{Value}\right\rangle$$

is called an epistemic act-utilitarian branching-time frame *(*eaubt*-frame for short) iff*

- $\left\langle M, \sqsubset, Ags, \mathbf{Choice}, \{\sim_\alpha\}_{\alpha\in Ags}, \mathbf{Value}\right\rangle$ *is a structure as in Definition 4.8, that additionally satisfies the following conditions:*
  - (`OAC`) Own action condition*: for all $\alpha \in Ags$ and each index $\langle m, h\rangle$, $\langle m, h\rangle \sim_\alpha \langle m, h'\rangle$ for every $h' \in \mathbf{Choice}_\alpha^m(h)$.*
  - (`Unif − H`) Uniformity of historical possibility*: for all $\alpha \in Ags$ and each index $\langle m, h\rangle$, if $\langle m, h\rangle \sim_\alpha \langle m', h'\rangle$, then for every $h_* \in H_m$ there exists $h'_* \in H_{m'}$ such that $\langle m, h_*\rangle \sim_\alpha \langle m', h'_*\rangle$.*

As a convention, I write $m \sim_\alpha m'$ if there exist $h \in H_m$ and $h' \in H_{m'}$ such that $\langle m, h\rangle \sim_\alpha \langle m', h'\rangle$.

An eaubt*-model $\mathcal{M}$ consists of the tuple that results from adding a valuation function $\mathcal{V}$ to an* eaubt*-frame, where $\mathcal{V} : P \to 2^{I(M\times H)}$ assigns to each atomic proposition a set of indices.*

As for (`OAC`) and (`Unif − H`), Broersen and Abarca (2018a) argued that these conditions are very useful, since together they imply the following properties for *EAUST*: (a) agents are able to knowingly do the same things across epistemically indistinguishable indices; (b) subjective ought-to-do's conform to Kant's directive of *ought implies can* in its subjective, epistemic version of *subjectively ought-to-do implies the ability to knowingly do*; and (c) if an agent subjectively ought to have done φ, then the agent must have known that this is the case. Indeed, the proof of soundness (Proposition C.40) for a proof system introduced in the next section certifies that conditions ($\mathtt{OAC}$) and ($\mathtt{Unif-H}$) are instrumental in granting points (a), (b), and (c).

Following Duijf (2018, Chapter 3), I refer to ($\mathtt{OAC}$) as *own action condition*. This frame condition entails that agents cannot know more than what their own actions bring about, and—as will be addressed in Subsection 4.5.1—it is defined by formula $K_\alpha\varphi \to [\alpha]\varphi$. Condition ($\mathtt{OAC}$) was introduced for the first time in Chapter 3, where I mentioned that it limits the analysis to cases in which an agent cannot have known what choices the other agents would perform (see Footnote 19, p. 93). In turn, and again following Duijf (2018, Chapter 3), I refer to ($\mathtt{Unif-H}$) as *uniformity of historical possibility*. This frame condition underlies a notion of uniformity of available actions, because it implies the following property in *eaubt*-models: at an index an agent was able in the epistemic sense to see to it that φ only if at the index the agent knew that it was possible to bring about φ. In other words, ($\mathtt{Unif-H}$) implies that agents are able to carry out the same actions at indistinguishable indices. As will also be addressed in Subsection 4.5.1, ($\mathtt{Unif-H}$) is defined by formula $\Diamond K_\alpha\varphi \to K_\alpha\Diamond\varphi$.

Eaubt-models allow us to provide semantics for the formulas of $\mathcal{L}_{\mathsf{KO}}$. However, before presenting the evaluation rules for such formulas, further definitions are required. The interesting cases concern modalities $\odot_\alpha\varphi$ and $\odot_\alpha^S\varphi$. As for $\odot_\alpha\varphi$, its semantics is the same as *AUST*'s, although I reinterpret $\odot_\alpha\varphi$ as a modality for objective ought-to-do's. As for $\odot_\alpha^S\varphi$, its semantics also results from a dominance ordering, but one that is different to the one used for objective ought-to-do's. To define this subjective dominance ordering, I make use of a new semantic concept, the *epistemic clusters* of a given choice of action:

Definition 4.19 (Epistemic clusters). *Let $\mathcal{M}$ be an* eaubt*-frame with M as its set of moments. Take $\alpha \in Ags$, and let $m, m' \in M$ be such that $m \sim_\alpha m'$. For $L \subseteq H_m$, L's* epistemic cluster at m' *is the set*

$$[L]_\alpha^{m'} := \{h' \in H_{m'};\ \textit{there is } h \in L \ \textit{s. t.}\ \ \langle m, h\rangle \sim_\alpha \langle m', h'\rangle\}.$$

Thus, for all $L \in \mathbf{Choice}_\alpha^m$, L's epistemic cluster at m' is nothing more than the set of histories in $H_{m'}$ anchoring indices that agent α cannot distinguish from those corresponding to L. As mentioned above, I use epistemic clusters to define a subjective dominance ordering over an agent's available actions at some moment.

In this subjective dominance ordering, the agent's choices are ranked taking into consideration its epistemic equivalents, so that the actions that are subjectively good enough will be the basis of said agent's subjective ought-to-do's:

Definition 4.20 (Subjective dominance ordering over choices). *For an* eaubt-*frame with M as its set of moments, $\alpha \in Ags$, and $m \in M$, consider the following definitions:*

- *Recall that $\leq$ is an ordering on 2^{H_m} such that $X \leq Y$ iff $\mathbf{Value}(h) \leq \mathbf{Value}(h')$ for every $h \in X$ and $h' \in Y$. Let $\leq_s$ be an ordering on $\mathbf{Choice}^m_\alpha$ defined by the rule: for $L, L' \in \mathbf{Choice}^m_\alpha$, $L \leq_s L'$ iff for all m' such that $m \sim_\alpha m'$ and each $S \in \mathbf{State}^{m'}_\alpha$, $[L]^{m'}_\alpha \cap S \leq [L']^{m'}_\alpha \cap S$. I write $L <_s L'$ iff $L \leq_s L'$ and $L' \not\leq_s L$. Observe that $L <_s L'$ iff (a) for all $m' \in M$ such that $m \sim_\alpha m'$ and each $S \in \mathbf{State}^{m'}_\alpha$, $[L]^{m'}_\alpha \cap S \leq [L']^{m'}_\alpha \cap S$, and (b) there exist $m_* \in M$ and $S_* \in \mathbf{State}^{m_*}_\alpha$ such that $m \sim_\alpha m_*$ and such that $[L']^{m_*}_\alpha \cap S \not\leq [L']^{m_*}_\alpha \cap S$.*
- *On the basis of $\leq_s$, a set of subjectively optimal actions is defined:* $\mathbf{SOptimal}^m_\alpha := \{L \in \mathbf{Choice}^m_\alpha;\ \text{there is no } L' \in \mathbf{Choice}^m_\alpha \text{ s. t. } L <_s L'\}$.

Observe that the definition of $\leq_s$ accounts for a form of sure-thing reasoning: for all $\alpha \in Ags$ and each moment m' such that $m \sim_\alpha m'$, the members of $\mathbf{State}^{m'}_\alpha$ provide a background against which α's actions are subjectively ranked. To clarify, once again consider Example 4.5 and its depiction in Figure 4.2. With the intuitively natural definition of $\sim_{Nik}$ given by $\langle m_1, h_1\rangle \sim_{Nik} \langle m_1, h_4\rangle$ and $\langle m_1, h_2\rangle \sim_{Nik} \langle m_1, h_2\rangle$, one has that $[H]^{m_1}_{Nik} = \{h_2, h_3\}$ and that $[T]^{m_1}_{Nik} = \{h_1, h_4\}$. Thus, for all $S \in \mathbf{State}^{m_1}_{Nik} = \{H(\text{for } \textit{Dolokhov}), T(\text{for } \textit{Dolokhov})\}$, $[T]^{m_1}_{Nik} \cap S \leq [H]^{m_1}_{Nik} \cap S$ and $[H]^{m_1}_{Nik} \cap S \not\leq [T]^{m_1}_{Nik} \cap S$, which implies that $H >_s T$. Thus, betting heads subjectively dominates betting tails for *Nikolai*, as desired.[21]

The subjective dominance ordering $\leq_s$ is used to define semantics for $\odot^S_\alpha\varphi$:

Definition 4.21 (Evaluation rules for *EAUST*). *Let $\mathcal{M}$ be an* eaubt-*model. The semantics on $\mathcal{M}$ for the formulas of $\mathcal{L}_{KO}$ are obtained by extending the recursive definition in Definition 4.9 with the following clause:*

$$\mathcal{M}, \langle m, h\rangle \models \odot^S_\alpha \varphi \quad \textit{iff} \quad \begin{array}{l} \textit{for all } L \in \mathbf{Choice}^m_\alpha \textit{ s. t. } \mathcal{M}, \langle m', h_L\rangle \not\models \varphi \textit{ for some } m' \textit{ s. t.} \\ m \sim_\alpha m' \textit{ and some } h_L \in [L]^{m'}_\alpha\textit{, there is } L' \in \mathbf{Choice}^m_\alpha \textit{ s. t.} \\ L <_s L' \textit{and, if } L'' = L' \textit{ or } L' \leq_s L''\textit{, then } \mathcal{M}, \langle m'', h''\rangle \models \varphi \\ \textit{for every } m'' \textit{ s. t. } m \sim_\alpha m'' \textit{ and every } h'' \in [L'']^{m''}_\alpha. \end{array}$$

Therefore, one says that at index $\langle m, h\rangle$ agent α subjectively ought to have seen to it that φ iff for each action L of α that does not guarantee the bringing about of φ

[21] It is worth noticing that Horty (2019) did not account for sure-thing reasoning in his logic for epistemic oughts. According to the definition of $\leq_H$ in Definition 4.14, $\mathit{Lbl}(H) \not\leq_H \mathit{Lbl}(T)$ in Example 4.5.

there is a subjectively better action L' such that (a) L's epistemic clusters guarantee the bringing about of φ, and (b) for each action that is subjectively better than L', all its epistemic clusters also guarantee the bringing about of φ. Just as in the case of objective ought-to-do's, such a truth condition can be much more intuitively stated when the number of available choices at each moment is finite. In this case, one has that

$$\begin{aligned}\mathcal{M},\langle m,h\rangle \models \odot^{S}_{\alpha}\varphi \quad \text{iff} \quad & \text{for all } L \in \textbf{SOptimal}^{m}_{\alpha}, \mathcal{M},\langle m',h'\rangle \models \varphi \\ & \text{for every } m' \text{ s. t. } m \sim_{\alpha} m' \text{ and every } h' \in [L]^{m''}_{\alpha}.\end{aligned}$$

The idea, then, is that at a given index α subjectively ought to have seen to it that φ iff φ is an effect of α's subjectively optimal actions at said index.

4.4.2 Solving the Problems in Horty's Puzzles

The semantics for subjective ought-to-do's allows us to solve the problems implied by Horty's puzzles. To be precise, consider Figures 4.3–4.5. For all three puzzles, the fact that *Dolokhov* hides his coin from *Nikolai* is captured by defining $\sim_{Nik}$ through the following information sets: $\{\langle m_2,h_1\rangle,\langle m_3,h_4\rangle\}$, in which *Nikolai* has bet heads; $\{\langle m_2,h_2\rangle,\langle m_3,h_5\rangle\}$, in which *Nikolai* has bet tails; and $\{\langle m_2,h_3\rangle,\langle m_3,h_6\rangle\}$, in which *Nikolai* has forfeited the bet.

For Puzzle #1 in Example 4.10, then, consider the *eaubt*-model $\mathcal{M}$ depicted in Figure 4.6.

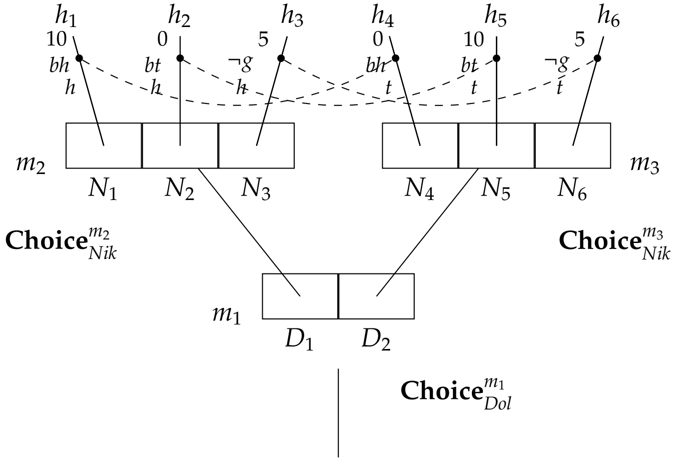


Figure 4.6: *Puzzle #1, revisited.*

Here, the setting coincides with Figure 4.3's in everything except *Nikolai*'s epistemic states, which are represented with the indistinguishability relation $\sim_{Nik}$ given by dashed lines (omitting reflexive loops). Thus, while in Horty's (2019) interpretation *Nikolai* could not distinguish between his available actions, in mine he could, certifying the fact that my models are more flexible than Horty's when it comes to indistinguishability relations.

Now, recall that Puzzle #1's problem was that $\mathcal{M}, \langle m_i, h\rangle \models K_{Nik} \odot_{Nik}\, g$ for all $i \in \{2, 3\}$ and $h \in H_{m_i}$: *at all indices based on moments* m_2 *and* m_3 Nikolai *knew that he ought to have gambled, even if gambling was a risky move that could result in a payoff of* 0. However, now I interpret this knowledge as the knowledge of an objective ought-to-do: *Nikolai* knew that he objectively ought to have gambled, in the sense that—objectively speaking—in every best choice he indeed has gambled. As for subjective ought-to-do's, however, one has that $\mathcal{M}, \langle m_i, h\rangle \models \neg\odot^{S}_{Nik} g$ for all $i \in \{2, 3\}$ and $h \in H_{m_i}$: *at all indices based on* m_2 *and* m_3 Nikolai *did not subjectively ought to have gambled*. This is a consequence of the following arguments:

- First, observe that $[N_1]^{m_2}_{Nik} = N_1$ and $[N_1]^{m_3}_{Nik} = N_4$, that $[N_2]^{m_2}_{Nik} = N_2$ and $[N_2]^{m_3}_{Nik} = N_5$, and that $[N_3]^{m_2}_{Nik} = N_3$ and $[N_3]^{m_3}_{Nik} = N_6$. Similarly, $[N_4]^{m_2}_{Nik} = N_1$ and $[N_4]^{m_3}_{Nik} = N_4$, $[N_5]^{m_2}_{Nik} = N_2$ and $[N_5]^{m_3}_{Nik} = N_5$, and $[N_6]^{m_2}_{Nik} = N_3$ and $[N_6]^{m_3}_{Nik} = N_6$.

- Secondly, observe that, for all $i \in \{2, 3\}$, every $S \in \mathbf{State}^{m_i}_{Nik}$ is such that $S = H_{m_i}$. Thus, the fact that $N_2 \le N_1$, which means that $[N_2]^{m_2}_{Nik} \le [N_1]^{m_2}_{Nik}$, implies that $N_2 \not\succ_s N_1$. Similarly, the fact that $N_3 \le N_1$ implies that $N_3 \not\succ_s N_1$. Now, the fact that $N_4 \le N_5$, which means that $[N_1]^{m_3}_{Nik} \le [N_2]^{m_3}_{Nik}$, implies that $N_1 \not\succ_s N_2$, and the fact that $N_6 \le N_5$, which means that $[N_3]^{m_3}_{Nik} \le [N_2]^{m_3}_{Nik}$, implies that $N_3 \not\succ_s N_2$. In turn, the fact that $N_4 \le N_6$, which means that $[N_1]^{m_3}_{Nik} \le [N_3]^{m_3}_{Nik}$, implies that $N_1 \not\succ_s N_3$, and the fact that $N_2 \le N_3$ implies that $N_2 \not\succ_s N_3$. Thus, $\mathbf{SOptimal}^{m_2}_{Nik} = \{N_1, N_2, N_3\}$. With analogous arguments, one can verify that $\mathbf{SOptimal}^{m_3}_{Nik} = \{N_4, N_5, N_6\}$.

- Thus, N_3 is such that $N_3 \in \mathbf{SOptimal}^{m_2}_{Nik}$ and $[N_3]^{m_2}_{Nik} = N_3 \subseteq |\neg g|^{m_2}$, and N_6 is such that $N_6 \in \mathbf{SOptimal}^{m_3}_{Nik}$ and $[N_6]^{m_3}_{Nik} = N_6 \subseteq |\neg g|^{m_3}$. Since the *eaubt*-model in Figure 4.6 is finite, one can use the evaluation rule for $\odot^{S}_{Nik}\varphi$ that is stated in terms of $\mathbf{SOptimal}^{m}_{\alpha}$ (p. 165). Therefore, N_3 and N_6 attest to the fact that $\mathcal{M}, \langle m_i, h\rangle \models \neg \odot^{S}_{Nik}\, g$ for all $i \in \{2, 3\}$ and $h \in H_{m_i}$, which is what we wanted to show.

Furthermore, Definition 4.21 also implies that $\mathcal{M}, \langle m_i, h\rangle \models K_{Nik}\neg \odot^{S}_{Nik}\, g$: *at all indices based on* m_2 *and* m_3 Nikolai *knew that he did not subjectively ought to have* *gambled*. Therefore, Puzzle #1's problem is solved, and the same solution applies for the variation of this example that represents the *Miners Paradox*. In this case, *rescuer* did not subjectively ought to have blocked a shaft.

For Puzzle #2 in Example 4.11, consider Figure 4.7.

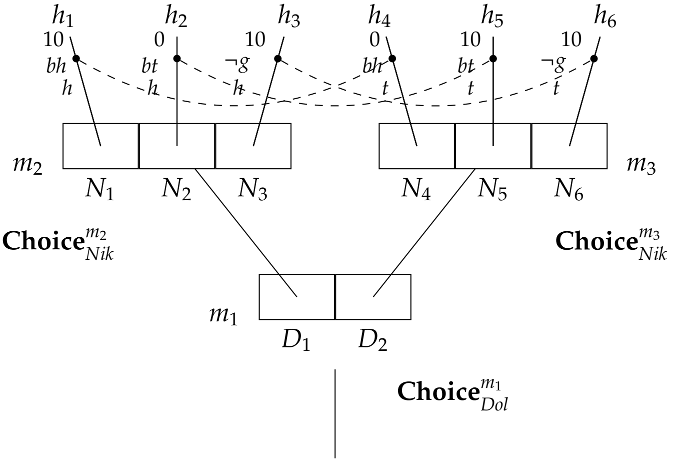


Figure 4.7: *Puzzle #2, revisited.*

Recall that Puzzle #2's problem was that $\mathcal{M},\langle m_i,h\rangle \not\models K_{Nik}\odot^S_{Nik}\neg g$ for all $i \in \{2,3\}$: *at all indices based on* m_2 *and* m_3 Nikolai *did not know that he subjectively ought to have not gambled*. However, now I interpret this as saying that *Nikolai* did not know that he objectively ought to have not gambled, which is a reasonable assumption on account of the fact that—objectively speaking—not every best choice involved not gambling. In contrast, reasoning about *Nikolai*'s subjective ought-to-do's leads us to conclude that $\mathcal{M},\langle m_i,h\rangle \models \odot^S_{Nik}\neg g$: *at all indices based on* m_2 *and* m_3 Nikolai *subjectively ought to have not gambled*. This is a consequence of the following arguments:

- The fact that $N_4 \leq N_6$ implies that $N_1 \not>_s N_3$, and the fact that $N_2 \leq N_3$ implies that $N_2 \not>_s N_3$. This time, however, observe that the fact that $N_6 \not\leq N_4$ implies that $N_3 >_s N_1$, and the fact that $N_3 \not\leq N_2$ implies that $N_3 >_s N_2$. Thus, $\mathbf{SOptimal}^{m_2}_{Nik} = \{N_3\}$, and, analogously, $\mathbf{SOptimal}^{m_3}_{Nik} = \{N_6\}$.
- Thus, for $N_3 \in \mathbf{SOptimal}^{m_2}_{Nik}$, $[N_3]^{m_2}_{Nik} = N_3 \subseteq |\neg g|^{m_2}$ and $[N_3]^{m_3}_{Nik} = N_6 \subseteq |\neg g|^{m_3}$. Similarly, for $N_6 \in \mathbf{SOptimal}^{m_3}_{Nik}$, $[N_6]^{m_2}_{Nik} = N_3 \subseteq |\neg g|^{m_2}$ and $[N_6]^{m_3}_{Nik} = N_6 \subseteq |\neg g|^{m_3}$. Definition 4.21 then implies that $\mathcal{M},\langle m_i,h\rangle \models \odot^S_{Nik}\neg g$ for all $i \in \{2,3\}$ and $h \in H_{m_i}$, which is what we wanted to show.

Furthermore, Definition 4.21 also implies that $\mathcal{M}, \langle m_i, h\rangle \models K_{Nik} \odot^{S}_{Nik} \neg g$ for all $i \in \{2,3\}$ and $h \in H_{m_i}$: *at all indices based on* m_2 *and* m_3 Nikolai *knew that he subjectively ought to have not gambled*. Therefore, Puzzle #2's problem is solved.

For Puzzle #3 in Example 4.12, consider Figure 4.8.

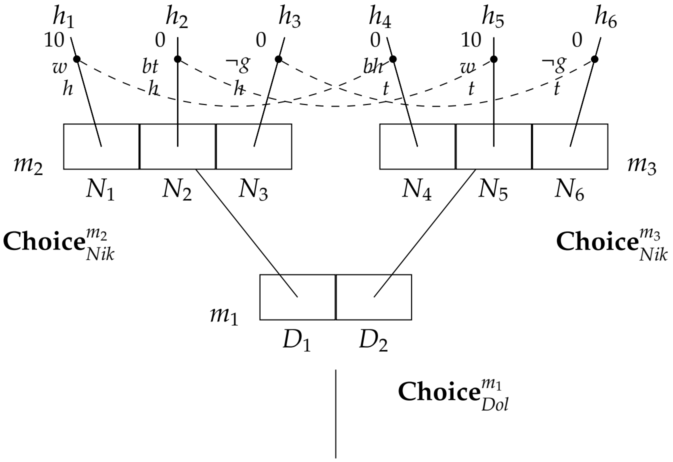


Figure 4.8: *Puzzle #3, revisited.*

Recall that Puzzle #3's problem was that $\mathcal{M}, \langle m_i, h\rangle \models K_{Nik} \odot_{Nik} w$ for all $i \in \{2,3\}$ and $h \in H_{m_i}$: *at all indices based on* m_2 *and* m_3 Nikolai *knew that he ought to have won*. Once again, now I interpret this statement as saying that *Nikolai* knew that he objectively ought to have won, which is a reasonable assumption on account of the fact that—objectively speaking—in every best choice he has won. In contrast, observe that $\mathcal{M}, \langle m_i, h\rangle \models \neg \odot^{S}_{Nik} w$ or $i \in \{2,3\}$ and $h \in H_{m_i}$: *at all indices based on* m_2 *and* m_3 Nikolai *did not subjectively ought to have won*. This is a consequence of the following arguments:

- The fact that $N_2 \leq N_1$ implies that $N_2 \not>_s N_1$, and the fact that $N_3 \leq N_1$ implies that $N_3 \not>_s N_1$. Similarly, the fact that $N_4 \leq N_5$ implies that $N_1 \not>_s N_2$, and the fact that $N_3 \leq N_2$ implies that $N_3 \not>_s N_2$. This time, however, the fact that $N_1 \not\leq N_3$ implies that $N_1 >_s N_3$. Thus, $\mathbf{SOptimal}^{m_2}_{Nik} = \{N_1, N_2\}$, and, analogously, $\mathbf{SOptimal}^{m_3}_{Nik} = \{N_4, N_5\}$.
- Thus, N_2 is such that $N_2 \in \mathbf{SOptimal}^{m_2}_{Nik}$ and $[N_2]^{m_2}_{Nik} = N_2 \subseteq |\neg w|^{m_2}$, and N_4 is such that $N_4 \in \mathbf{SOptimal}^{m_3}_{Nik}$ and $[N_4]^{m_3}_{Nik} = N_4 \subseteq |\neg w|^{m_3}$. Definition 4.21 then implies that $\mathcal{M}, \langle m_i, h\rangle \models \neg \odot^{S}_{Nik} w$ for all $i \in \{2,3\}$ and $h \in H_{m_i}$, which is what we wanted to show.

Furthermore, Definition 4.21 implies that $\mathcal{M}, \langle m_i, h \rangle \models K_{Nik} \neg \odot^{\mathcal{S}}_{Nik}\, w$ for all $i \in \{2, 3\}$ and $h \in H_{m_i}$: *at all indices based on* m_2 *and* m_3 Nikolai *knew that he did not subjectively ought to have won*. Therefore, Puzzle #3's problem is solved.[22]

4.4.3 Relation to Horty's Framework of Epistemic Oughts

When comparing my solution to the puzzles' problems with Horty's (2019), it is important to point out that my formalism is different from his in four main points:

1. In *eaubt*-models, indistinguishability relations occur at the level of indices, while—as mentioned in Subsection 4.3.1 (item 1 on p. 160)—Horty's models include indistinguishability relations at the level of moments.
2. *Eaubt*-models do not include action types.
3. *EAUST* is not restricted to finite-choice models, meaning that agents' choice-partitions can also be infinite.
4. The dominance ordering for my subjective ought-to-do's accounts for a form of sure-thing reasoning that is absent in Horty's treatment of epistemic oughts (see Footnote 21).

Regardless of these differences, the respective solutions are virtually the same, as represented in Table 4.1.

Solution / *Puzzle*	Horty's	Mine
Ex. 4.10 (Puzzle #1)	$K_{Nik} \neg \odot [Nik\ \texttt{kstit}] g$	$K_{Nik} \neg \odot^{\mathcal{S}}_{Nik}\, g$
Ex. 4.11 (Puzzle #2)	$K_{Nik} \odot [Nik\ \texttt{kstit}] \neg g$	$K_{Nik} \odot^{\mathcal{S}}_{Nik} \neg g$
Ex. 4.12 (Puzzle #3)	$K_{Nik} \neg \odot [Nik\ \texttt{kstit}] w$	$K_{Nik} \neg \odot^{\mathcal{S}}_{Nik}\, w$

Table 4.1: *Comparison of solutions.*

Therefore, $\odot^{\mathcal{S}}_{\alpha}$ works as an analog of $\odot[\alpha\ \texttt{kstit}]$.

Besides the fact that both approaches solve the puzzles' problems, there is a stronger connection between the framework with action types (Horty, 2019; Horty & Pacuit, 2017) and the one developed here. Let me elaborate on this connection.

[22] Interestingly, observe that, in this case, $\mathcal{M}, \langle m_i, h \rangle \models \odot^{\mathcal{S}}_{Nik} g$: *at all indices based on* m_2 *and* m_3 *Nikolai subjectively ought to have gambled.*

Duijf et al. (2021) established a correspondence result between Horty-like labelled *ebt*-models (which are those restrictions of Definition 4.14's models that do not include **Value** and its associated notions), on the one hand, and *ebt*-models (Definition 2.28, p. 70), on the other. As shown below, this result can be extended to formulas of the language $\mathcal{L}_{\mathsf{H}}$, the language that includes the deontic operators $\odot[\alpha\ \texttt{stit}]$ and $\odot[\alpha\ \texttt{kstit}]$.

The basic idea is to first define a translation Tr from $\mathcal{L}_{\mathsf{H}}$ to $\mathcal{L}_{\mathsf{KO}}$. One then shows that a finite-choice Horty-like labelled *eaubt*-model $\mathcal{M}$ can be used to build a finite-choice *eaubt*-model $\mathcal{M}'$ such that, for every formula φ of $\mathcal{L}_{\mathsf{H}}$, φ holds at an index in $\mathcal{M}$ iff $Tr(\varphi)$ holds at the same index in $\mathcal{M}'$, provided that one weakens the definition of $\preceq_s$ in $\mathcal{M}'$ and drops sure-thing reasoning. Following Duijf et al. (2021), I refer to $\mathcal{M}'$ as the *transform structure* of $\mathcal{M}$. I address the basics of such a correspondence result below.

Definition 4.22 (Translation). *Assume that both $\mathcal{L}_H$ and $\mathcal{L}_{KO}$ are based on the same set P of propositional letters and on the same set Ags of agent names. A translation function $Tr : \mathcal{L}_H \to \mathcal{L}_{KO}$ is recursively defined by setting*

$$\begin{array}{lcl}
Tr(p) & = & p \\
Tr(\neg\varphi) & = & \neg Tr(\varphi) \\
Tr(\varphi \wedge \psi) & = & Tr(\varphi) \wedge Tr(\psi) \\
Tr(\Box\varphi) & = & \Box Tr(\varphi) \\
Tr([\alpha\ \texttt{stit}]\varphi) & = & [\alpha]Tr(\varphi) \\
Tr(K_\alpha\varphi) & = & \Box K_\alpha Tr(\varphi) \\
Tr([\alpha\ \texttt{kstit}]\varphi) & = & K_\alpha Tr(\varphi) \\
Tr(\odot[\alpha\ \texttt{stit}]\varphi) & = & \odot_\alpha Tr(\varphi) \\
Tr(\odot[\alpha\ \texttt{kstit}]\varphi) & = & \odot_\alpha^S Tr(\varphi).
\end{array}$$

Definition 4.23 (Transform structure). *Let $\mathcal{F}$ be a labelled* eaubt*-frame of the form $\left\langle M, \sqsubset, Ags, \mathbf{Choice}, \{\sim_\alpha\}_{\alpha\in Ags}, \mathbf{Value}, Tps, Lbl, Exe\right\rangle$. The tuple $\mathcal{F}' := \left\langle M', \sqsubset', Ags', \mathbf{Choice}', \{\sim'_\alpha\}_{\alpha\in Ags}, \mathbf{Value}'\right\rangle$ is called the* transform structure of *$\mathcal{F}$ iff $M' = M$, $\sqsubset' = \sqsubset$, $Ags' = Ags$, $\mathbf{Choice}' = \mathbf{Choice}$, $\mathbf{Value}' = \mathbf{Value}$, and, for $\alpha \in Ags$, $\sim'_\alpha$ is defined on $I(M \times H)$ by the following rule: for indices $\langle m, h\rangle$ and $\langle m', h'\rangle$, $\langle m, h\rangle \sim'_\alpha \langle m', h'\rangle$ iff $\langle m, h\rangle \sim_\alpha \langle m', h'\rangle$ and $Lbl_\alpha(\langle m, h\rangle) = Lbl_\alpha(\langle m', h'\rangle)$.*

For a labelled eaubt*-model $\mathcal{M}$ that is based on labelled* eaubt*-frame $\mathcal{F}$ and that has valuation function $\mathcal{V}$, the* eaubt*-model $\mathcal{M}'$ that results from adding $\mathcal{V}$ to $\mathcal{F}'$—the transform structure of $\mathcal{F}$—is known as the* transform structure of *$\mathcal{M}$.*

The following proposition, whose proof is relegated to Appendix B, guarantees that the transform structures of labelled *eaubt*-frames are indeed *eaubt*-frames.

Proposition 4.24. *Let $\mathcal{F}$ be a labelled* eaubt*-frame, and let $\mathcal{F}'$ be its transform structure. Then $\mathcal{F}'$ is an unconstrained* eaubt*-frame. Additionally, if $\mathcal{F}$ is Horty-like, then $\mathcal{F}'$ is an* eaubt*-frame, and if $\mathcal{F}$ is finite-choice, then $\mathcal{F}'$ is also finite-choice.*

Let me illustrate these transform structures. Recall that, when presenting Horty's solution to the puzzles of Section 4.3, the strategy was to interpret Figures 4.3, 4.4, and 4.5 as Horty-like labelled *eaubt*-models. My interpretations of Horty's puzzles, discussed in Subsection 4.4.2, are actually the transform *eaubt*-models of the Horty-like labelled *eaubt*-models in Figures 4.3, 4.4, and 4.5. To be precise, the *eaubt*-model depicted in Figure 4.6—resp. 4.7, resp. 4.8—is the transform *eaubt*-model of the Horty-like labelled *eaubt*-model of Figure 4.3—resp. 4.4, resp. 4.5.

The correspondence result, then, is given by the theorem below, whose proof is relegated to Appendix B.

Theorem 4.25 (Correspondence). *Let $\mathcal{M}$ be a finite-choice Horty-like labelled* eaubt*-model, and let $\mathcal{M}'$ be its transform finite-choice* eaubt*-model. Let us redefine $\leq_s$ in $\mathcal{M}'$ so that, for $\alpha \in Ags$, $m \in M$, and $L, L' \in$* **Choice**$^m_\alpha$*, $L \leq_s L'$ iff for all m' such that $m \sim_\alpha m'$, $[L]^{m'}_\alpha \leq [L']^{m'}_\alpha$. Then, for every formula φ of $\mathcal{L}_{\mathsf{H}}$, $\mathcal{M}, \langle m, h\rangle \models \varphi$ iff $\mathcal{M}', \langle m, h\rangle \models Tr(\varphi)$.*

Thus, finite-choice Horty-like labelled *eaubt*-models correspond to a sub-class of *eaubt*-models, namely finite-choice *eaubt*-models. Proposition 4.24 and Theorem 4.25 imply that, while one can simulate Horty's (2019) logic of epistemic oughts using mine, a converse simulation can only hold if Horty's logic is adapted to handle cases with infinite choices.

4.5 Logic-Based Properties & Axiomatization

4.5.1 Properties

Let me present and discuss interesting properties of *EAUST*, in terms of formulas that are either valid or invalid with respect to *eaubt*-models. As for the logic-based properties of modalities $\Box\varphi$ and $[\alpha]\varphi$, they are the same as the ones reviewed in Chapter 2's Subsection 2.3.1: both operators are **S5**, and they validate the schemata known as (*SET*) and (*IA*). For every $\alpha \in Ags$, K_α is also **S5**. Thus, the properties of knowledge are the ones reviewed in Chapter 2 Subsection 2.4.4: logical omniscience, factivity, positive introspection, and negative introspection.

As for the deontic modalities, it turns out that both $\odot_\alpha$ and $\odot^S_\alpha$ are **KD45** operators. The validity of schema (*K*) entails that the logical consequences of obligations are also obligations; the validity of schema (*D*) ($\Delta\varphi \to \neg\Delta\neg\varphi$ for $\Delta \in$

$\left\{\odot_\alpha, \odot^S_\alpha\right\}$) entails that both objective and subjective ought-to-do's are respectively consistent; the validity of schema (4) ($\Delta\varphi \to \Delta\Delta\varphi$ for $\Delta \in \left\{\odot_\alpha, \odot^S_\alpha\right\}$) entails that, for both objective and subjective senses, if at an index an agent had an obligation, then the agent also ought to have seen to it that the agent itself had that obligation; and the validity of schema (5) ($\neg\Delta\varphi \to \Delta\neg\Delta\varphi$ for $\Delta \in \left\{\odot_\alpha, \odot^S_\alpha\right\}$) entails that, for both objective and subjective senses, if at an index an agent did not have an obligation, then the agent ought to have seen to it that the agent itself did not have that obligation. According to Horty's (2001, Chapter 4) seminal theory of ought-to-do, all these are reasonable and desirable properties of obligations.

Furthermore, the validity, resp. invalidity, of the following formulas, with respect to the class of *eaubt*-models, captures desirable properties for the interplay between the modalities of *EAUST*.

1. (a) $\models \odot_\alpha\varphi \to \Box \odot_\alpha \varphi$: if at an index an agent objectively ought to have seen to it that φ, then this was settled at the index.

   (b) $\models \neg \odot_\alpha \varphi \to \Box\neg \odot_\alpha \varphi$: if at an index an agent did not objectively ought to have seen to it that φ, then this was settled at the index.

   With properties 2–4 below, these two are standard in *AUST* (Horty, 2001; Murakami, 2004). A proof of validity—for both formulas—follows from the truth condition for $\odot_\alpha\varphi$.[23]

2. $\models \Box\varphi \to \odot_\alpha\varphi$: if at an index φ was settled, then every agent objectively ought to have seen to it that φ. A proof of validity follows from the truth condition for $\odot_\alpha\varphi$.

3. $\models \odot_\alpha\varphi \to \Diamond[\alpha]\varphi$: if at an index an agent objectively ought to have seen to it that φ, then it must have been historically possible for that agent to see to it that φ. This is the objective version of Kant's directive of *ought implies can* (see Horty, 2001, Chapter 4), so that an agent objectively ought to have brought about φ only if it was causally able to bring about φ.

4. $\not\models \odot_\alpha\varphi \to [\alpha]\varphi$: it is not necessarily true that if at an index an agent ought to have seen to it that φ then that agent has seen to it that φ. The models used for the examples in this chapter all offer counterexamples. For instance, in Figure 4.6, $\langle m_2, h_2\rangle$ is such that $\mathcal{M}, \langle m_2, h_2\rangle \models \odot_\alpha bh$ and $\mathcal{M}, \langle m_2, h_2\rangle \not\models [\alpha]bh$.

5. (a) $\models K_\alpha\varphi \to [\alpha]\varphi$: if at an index an agent knew φ, then the agent has actually seen to it that φ. In the discussion after Definition 4.18 (p. 162), I mentioned that the validity of this formula is associated with frame

[23] In fact, the second formula can be derived using the first, as shown in Proposition C.38.

condition (OAC) and that it implies that an agent cannot know more than what it brings about. Moreover, $K_\alpha\varphi \to [\alpha]\varphi$ defines (OAC) (see Blackburn et al., 2002, Chapter 3, for the precise definitions of frame definability through modal formulas), since it is easy to see that an unconstrained *eaubt*-frame satisfies (OAC) iff $K_\alpha\varphi \to [\alpha]\varphi$ is valid on said frame. For a proof of validity, see the (OAC) item in Proposition C.40.

(b) $\models K_\alpha\varphi \leftrightarrow K_\alpha[\alpha]\varphi$: to know φ is the same as to knowingly see to it that φ. This formula characterizes knowledge in *EAUST*, so that, according to my treatment of knowledge across the stages of information disclosure (p. 152), $K_\alpha\varphi$ expresses that agent α had *ex interim* knowledge of φ. Furthermore, since the validity of $K_\alpha\varphi \leftrightarrow K_\alpha[\alpha]\varphi$ implies the validity of $\Box K_\alpha\varphi \leftrightarrow \Box K_\alpha[\alpha]\varphi$, then *ex ante* knowledge boils down to *ex interim* knowledge that holds regardless of anyone's choice of action. For a proof of validity of $K_\alpha\varphi \leftrightarrow K_\alpha[\alpha]\varphi$, observe that it is implied by the validity of $K_\alpha\varphi \to [\alpha]\varphi$ (item 5a above), coupled with the facts that schemata (K), (T), and (4) for K_α are valid.

6. (a) $\models \Diamond K_\alpha\varphi \to K_\alpha\Diamond\varphi$: if at an index it was historically possible for an agent to know φ, then the agent knew that φ was historically possible at the index. As also mentioned in the discussion after Definition 4.18, the validity of this formula is associated with frame condition (Unif − H). Indeed, it defines (Unif − H) insofar as an unconstrained *eaubt*-frame satisfies (Unif − H) iff $\Diamond K_\alpha\varphi \to K_\alpha\Diamond\varphi$ is valid on said frame. For a proof of validity, see the ($Unif - H$) item in Proposition C.40.

(b) $\models \Diamond K_\alpha[\alpha]\varphi \to K_\alpha\Diamond[\alpha]\varphi$: if at an index it was historically possible for an agent to knowingly see to it that φ, then the agent knew that it was possible to see to it that φ at the index. This formula encodes a requirement of uniformity, according to which agents should be able to carry out the same actions at indistinguishable indices. Thus, one of the essential targets for my notion of subjective ought-to-do's—point (a) in the discussion after Definition 4.18—has been met. In light of item 5b above, it is easy to see how the validity of this formula is equivalent to that of $\Diamond K_\alpha\varphi \to K_\alpha\Diamond\varphi$.[24]

(c) $\models \Box K_\alpha\varphi \leftrightarrow K_\alpha\Box\varphi$: at an index an agent knew that φ was settled iff it was settled that the agent knew φ. This formula characterizes *ex*

[24] Duijf et al.'s (2021) method can be adapted to show how $\Diamond K_\alpha\varphi \to K_\alpha\Diamond\varphi$ (and thus $\Diamond K_\alpha[\alpha]\varphi \to K_\alpha\Diamond[\alpha]\varphi$) correspond to the constraint of *uniformity of available action types* (UAAT) that Horty (2019) includes in his Horty-like labelled *eaubt*-models. Observation 4.31 a shows that, even without the validity of the formula in item 5b, the validity of $\Diamond K_\alpha[\alpha]\varphi \to K_\alpha\Diamond[\alpha]\varphi$ is equivalent to the validity of $\Diamond K_\alpha\varphi \to K_\alpha\Diamond\varphi$.

ante knowledge in *EAUST*, so that at a given index an agent had *ex ante* knowledge of φ iff the agent knew that φ was settled. The validity of this formula is also equivalent to that of $\Diamond K_\alpha\varphi \to K_\alpha\Diamond\varphi$ (Observation 4.31 b).

7. $\models \odot^S_\alpha\varphi \to \odot^S_\alpha (K_\alpha[\alpha]\varphi)$: if at an index an agent subjectively ought to have seen to it that φ, then the agent subjectively ought to have brought about that itself has knowingly seen to it that φ. In other words, subjective ought-to-do's concern states of affairs that an agent not only should bring about, but that it should bring about knowingly. This property carries one step further the natural correlation between subjective ought-to-do's and knowingly doing, just as planned by Horty (2019).[25] Given the validity of the formulas associated with (`OAC`) (item 5 above), this formula is equivalent to $\odot^S_\alpha\varphi \to \odot^S_\alpha(K_\alpha\varphi)$, which is schema ($A6$) in the proof system for the logic of subjective ought-to-do's in Definition 4.29. Thus, for a proof of validity, see the ($A6$) item in Proposition C.40.

8. $\not\models \odot_\alpha\varphi \to \Diamond K_\alpha[\alpha]\varphi$: it is not necessarily true that if at an index an agent objectively ought to have seen to it that φ then the agent could have knowingly seen to it that φ. Figure 4.8 offers a counterexample, because for all $i \in \{2,3\}$ and $h \in H_{m_i}$ $\mathcal{M}, \langle m_i, h\rangle \models \odot_{Nik}w$ and $\mathcal{M}, \langle m_i, h\rangle \not\models \Diamond K_{Nik}[Nik]w$: *at all indices based on* m_2 *and* m_3 Nikolai *objectively ought to have won, but it was impossible for him to knowingly win* (as witnessed by the facts that $\mathcal{M}, \langle m_3, h_4\rangle \not\models [Nik]w$, that $\mathcal{M}, \langle m_2, h_2\rangle \not\models [Nik]w$, and that $\mathcal{M}, \langle m_2, h_3\rangle \not\models [Nik]w$).

9. $\models \odot^S_\alpha\varphi \to \Diamond K_\alpha\varphi$: if at an index an agent subjectively ought to have seen to it that φ, then it must have been possible for the agent to know φ. Given the validity of the formulas associated with (`OAC`) in item 5, the validity of this formula implies that of $\odot^S_\alpha\varphi \to \Diamond K_\alpha[\alpha]\varphi$, which is the subjective version of Kant's directive of *ought implies can*: if at an index an agent subjectively ought to have seen to it that φ, then it must have been possible for the agent to knowingly see to it that φ. Thus, this property means that another essential target for my notion of subjective ought-to-do's—point (b) in the discussion after Definition 4.18—has been met. For a proof of validity, see the (*s.Oic*) item in Proposition C.40.

[25] As mentioned right before Definition 4.17 (p. 162), objective ought-to-do's are correlated with causal agency, and subjective ought-to-do's are correlated with epistemic agency. Observe, then, that formula $\odot_\alpha\varphi \to \odot_\alpha([\alpha]\varphi)$ is indeed valid in the case of objective ought-to-do's, ratifying such a correlation (Horty, 2001; Murakami, 2004) (see also schema ($A4$) in Definition 4.27).

10. $\not\models \odot_\alpha \varphi \to K_\alpha \odot_\alpha \varphi$: it is not necessarily true that if at an index an agent objectively ought to have seen to it that φ then the agent knew about such an objective obligation. Figure 4.8 offers a counterexample, because $\mathcal{M}, \langle m_2, h_1 \rangle \models \odot_{Nik} bh$ and $\mathcal{M}, \langle m_2, h_1 \rangle \not\models K_{Nik} \odot_{Nik} bh$: *at* $\langle m_2, h_1 \rangle$ Nikolai *objectively ought to have bet heads, but he did not know this* (as witnessed by the fact that $\mathcal{M}, \langle m_3, h_4 \rangle \models \odot_{Nik} bt$).

11. (a) $\models \odot^S_\alpha \varphi \to K_\alpha \Box \odot^S_\alpha \varphi$: if at an index an agent subjectively ought to have seen to it that φ, then the agent knew that such a subjective obligation was settled at that index. Given the validity of $\Box K_\alpha \varphi \leftrightarrow K_\alpha \Box \varphi$ (item 6c in this list), this formula implies that an agent's subjective obligations are always known *ex ante* by the agent. Thus, this property means that the last of the essential targets for my notion of subjective ought-to-do's—point (c) in the discussion after Definition 4.18—has been met. For a proof of validity, see the (*s.Cl*) item in Proposition C.40.

    (b) $\models \neg \odot^S_\alpha \varphi \to K_\alpha \Box \neg \odot^S_\alpha \varphi$: if at an index an agent did not subjectively ought to have seen to it that φ, then the agent knew that this lack of a subjective obligation was settled at that index. In the proof system for the logic of subjective ought-to-do's given in Definition 4.29, this formula can be derived using $\odot^S_\alpha \varphi \to K_\alpha \Box \odot^S_\alpha \varphi$, so it is also valid (Observation 4.31 c).

12. (a) $\not\models \odot^S_\alpha \varphi \to \odot_\alpha \varphi$: it is not necessarily true that if at an index an agent subjectively ought to have seen to it that φ then the agent objectively ought to have seen to it that φ. Figure 4.7 offers a counterexample, because $\mathcal{M}, \langle m_2, h_1 \rangle \models \odot^S_{Nik} \neg g$ and $\mathcal{M}, \langle m_2, h_1 \rangle \not\models \odot_{Nik} \neg g$: *at* $\langle m_2, h_1 \rangle$ Nikolai *subjectively ought to have not gambled, but he was not objectively obligated to not gamble.*

    (b) $\not\models \odot_\alpha \varphi \to \odot^S_\alpha \varphi$: it is not necessarily true that if at an index an agent objectively ought to have seen to it that φ then the agent subjectively ought to have seen to it that φ. Figure 4.8 offers a counterexample, because for all $i \in \{2, 3\}$ and $h \in H_{m_i}$ $\mathcal{M}, \langle m_i, h \rangle \models \odot_{Nik} w$ and $\mathcal{M}, \langle m_i, h \rangle \not\models \odot^S_{Nik} w$: *at all indices based on* m_2 *and* m_3 Nikolai *objectively ought to have won, but he was to subjectively obligated to win.*

    Thus, it turns out that modality $\odot^S_\alpha \varphi$ is neither (logically) stronger nor weaker than $\odot_\alpha \varphi$, just as Horty's (2019) modality for epistemic oughts is neither stronger nor weaker than $\odot_\alpha \varphi$.

13. $\models \odot_\alpha \varphi \to \neg \odot^S_\alpha \neg \varphi$ and $\models \odot^S_\alpha \varphi \to \neg \odot_\alpha \neg \varphi$: for each agent, its objective and subjective ought-to-do's are consistent. In other words, an agent cannot

have been objectively, resp. subjectively, obligated to see to it that φ and at the same time subjectively, resp. objectively, obligated to see to it that $\neg\varphi$. This property reflects a desirable tenet of consistency between senses of obligation that are both based on dominance of choices.[26] A proof of validity is included in Observation C.39.

4.5.2 Axiomatization

In this subsection I introduce two proof systems, one for the logic of objective ought-to-do's, and one for the logic of subjective ought-to-do's. To clarify, consider the following disambiguation: the logic of objective ought-to-do's has language $\mathcal{L}_{\mathsf{O}}$ (Definition 4.2), and the formulas are evaluated on *aubt*-models (Definition 4.3); the logic of subjective ought-to-do's has a language $\mathcal{L}_{\mathsf{S}}$, that is defined as a restriction of $\mathcal{L}_{\mathsf{KO}}$ as follows:

Definition 4.26 (Syntax for the logic of subjective ought-to-do's)**.** *The grammar for the formal language $\mathcal{L}_{\mathsf{S}}$ is given by*

$$\varphi ::= p \mid \neg\varphi \mid \varphi \wedge \varphi \mid \Box\varphi \mid [\alpha]\varphi \mid K_\alpha\varphi \mid \odot^{S}_{\alpha}\varphi,$$

where p ranges over P and α ranges over Ags.

The modalities of this language are the same as those in Definition 4.17. As for the semantics, the formulas of $\mathcal{L}_{\mathsf{S}}$ are evaluated on *eaubt*-models (Definition 4.18).

In what follows, I address the soundness & completeness results for both systems independently, and only afterwards discuss a joint proof system and its metalogic properties. As for objective ought-to-do's, their logic is axiomatized according to Definition 4.27 and Theorem 4.28 below.

Definition 4.27 (Proof system for objective ought-to-do's)**.** *Let Λ_{O} be the proof system defined by the following axioms and rules of inference:*

[26] For a discussion of another sense of obligation—based on maximization of expected utility—that is inconsistent both with the objective sense and with the subjective one, see this chapter's conclusion (Subsection 4.6.1).

- (Axioms) *All classical tautologies from propositional logic; the* **S5** *schemata for* $\Box$ *and* $[\alpha]$*; and the following schemata:*

$$
\begin{array}{ll}
\Box\varphi \to [\alpha]\varphi & (SET) \\
\textit{For all } n \geq 1 \textit{ and pairwise different } \alpha_1, \dots, \alpha_n, & \\
\bigwedge_{1 \leq k \leq n} \Diamond[\alpha_i]\varphi_i \to \Diamond\left(\bigwedge_{1 \leq k \leq n}[\alpha_i]\varphi_i\right) & (IA) \\
\odot_\alpha(\varphi \to \psi) \to (\odot_\alpha\varphi \to \odot_\alpha\psi) & (A1) \\
\Box\varphi \to \odot_\alpha\varphi & (A2) \\
\odot_\alpha\varphi \to \Box \odot_\alpha \varphi & (A3) \\
\odot_\alpha\varphi \to \odot_\alpha([\alpha]\varphi) & (A4) \\
\odot_\alpha\varphi \to \Diamond[\alpha]\varphi & (Oic)
\end{array}
$$

- *(Rules of inference) Modus Ponens, Substitution, and Necessitation for all modal operators.*

Schemata (SET) and (IA) are standard in *BST*, and they were discussed in Chapter 2's Subsection 2.3.1. Schema $(A1)$ ensures that $\odot_\alpha$ is a normal modal operator. Schema $(A2)$ characterizes syntactically that if φ was settled at an index then every agent ought to have seen to it that φ at that index. Schema $(A3)$ characterizes syntactically that if at an index an agent ought to have seen to it that φ then such an obligation was settled at the index. Schema $(A4)$ characterizes syntactically that if at an index an agent ought to have seen to it that φ then the agent ought to have seen to it that the agent itself has seen to it that φ. Schema (Oic)—where 'Oic' stands for *ought implies can*—concerns Kant's directive of *ought implies can* in the objective sense: if at an index an agent ought to have seen to it that φ, then it must have been historically possible for the agent to see to it that φ at the index.

Now, Murakami (2004) gave a slightly different proof system for this logic of objective ought-to-do's, and she showed that hers is sound and complete with respect to the class of *aubt*-models, as well as decidable. Thus, soundness and completeness of Λ_O are direct consequences of the equivalence between Murakami's (2004) proof system and Λ_O. Such an equivalence is proved in Appendix C (Proposition C.38), yielding the following metalogic result.

Theorem 4.28 (Soundness & Completeness of Λ_O)**.** *The proof system* Λ_O *is sound and complete with respect to the class of* aubt*-models.*

As for subjective ought-to-do's, their logic is axiomatized according to Definition 4.29 and Theorem 4.30 below.

Definition 4.29 (Proof system for subjective ought-to-do's)**.** *Let* Λ_S *be the proof system defined by the following axioms and rules of inference:*

- (Axioms) *All classical tautologies from propositional logic; the* **S5** *schemata for* $\Box$, $[\alpha]$, *and* K_α*; and the following schemata for the interactions between modalities:*

$$
\begin{array}{ll}
\Box\varphi \to [\alpha]\varphi & (SET) \\
\textit{For all } m \geq 1 \textit{ and pairwise different } \alpha_1, \ldots, \alpha_m, & \\
\bigwedge_{1\leq k\leq m} \Diamond[\alpha_i]\varphi_i \to \Diamond\left(\bigwedge_{1\leq k\leq m}[\alpha_i]\varphi_i\right) & (IA) \\
K_\alpha\varphi \to [\alpha]\varphi & (OAC) \\
\Diamond K_\alpha\varphi \to K_\alpha\Diamond\varphi & (Unif-H) \\
\odot^S_\alpha(\varphi \to \psi) \to (\odot^S_\alpha\varphi \to \odot^S_\alpha\psi) & (A5) \\
\odot^S_\alpha\varphi \to \odot^S_\alpha(K_\alpha\varphi) & (A6) \\
K_\alpha\Box\varphi \to \odot^S_\alpha\varphi & (SuN) \\
\odot^S_\alpha\varphi \to \Diamond K_\alpha\varphi & (s.Oic) \\
\odot^S_\alpha\varphi \to K_\alpha\Box\odot^S_\alpha\varphi & (s.Cl)
\end{array}
$$

- *(Rules of inference) Modus Ponens, Substitution, and Necessitation for all modal operators.*

Schemata (*SET*) and (*IA*) characterize the same properties as in the logic of objective ought-to-do's, but in *eaubt*-models.

Schema (*OAC*)—where 'OAC' stands for *own action condition*—encodes frame condition (OAC) (see Definition 4.18 and item 5 in the list of logic-based properties from the previous section). Schema (*Unif* − *H*)—where 'Unif-H' stands for *uniformity of historical possibility*—characterizes syntactically frame condition (Unif − H) (see Definition 4.18 and item 6 in the list of logic-based properties).

Schema (*A*5) ensures that $\odot^S_\alpha$ is a normal modal operator. Schema (*A*6) characterizes syntactically that if at an index an agent subjectively ought to have seen to it that φ then the agent ought to have seen to it that the agent itself knew φ (see item 7 in the list of logic-based properties). Schema (*SuN*)—where 'SuN' stands for *subjective necessity*—characterizes syntactically that if at an index an agent knew that φ was historically necessary then the agent must have been subjectively obligated to bring about φ. In other words, an agent's *ex ante* knowledge of φ implies that the agent subjectively ought to have seen to it that φ. Schema (*s.Oic*)—where 's.Oic' stands for *subjective ought implies can*—concerns the subjective version of Kant's directive of *ought implies can* (see item 9 in the list of logic-based properties). Finally, schema (*s.Cl*)—where 's.Cl' stands for *subjective closure*—characterizes syntactically that if at an index an agent subjectively ought to have seen to it that φ then the agent knew that this was settled (see item 11 in the list of logic-based properties).

Theorem 4.30 (Soundness & Completeness of Λ_S)**.** *The proof system Λ_S is sound and complete with respect to the class of* eaubt-*models.*

The proof of Theorem 4.30 is the main technical contribution of this chapter, and it is relegated to Appendix C. As for soundness, the proof is standard. As for completeness, the proof is a two-step process. First, I introduce a Kripke semantics for the logic, where the formulas of $\mathcal{L}_S$ are evaluated on Kripke-*eaus*-models (Definition C.41). I prove completeness of Λ_S with respect to the these structures, via the well-known technique of canonical models. Secondly, a truth-preserving correspondence between Kripke-*eaus*-models and a sub-class of *eaubt*-models is used for proving completeness with respect to *eaubt*-models via completeness with respect to Kripke-*eaus*-models.

Observation 4.31. *To illustrate the derivation of theorems in Λ_S, consider the following Λ_S-theorems, which are all important according to the list of logic-based properties given in the previous section:*

(a) $\Diamond K_\alpha[\alpha]\varphi \rightarrow K_\alpha\Diamond[\alpha]\varphi$. It is obtained by Substitution on schema ($Unif - H$), substituting $[\alpha]\varphi$ for φ. Interestingly, substituting $\Diamond K_\alpha[\alpha]\varphi \rightarrow K_\alpha\Diamond[\alpha]\varphi$ for ($Unif - H$) in Λ_S yields the same theory, since one can obtain $\Diamond K_\alpha\varphi \rightarrow K_\alpha\Diamond\varphi$ from $\Diamond K_\alpha[\alpha]\varphi \rightarrow K_\alpha\Diamond[\alpha]\varphi$ according to the following derivation, where 'c.p.' abbreviates 'contrapositive,' 'Nec.' abbreviates 'Necessitation,' and 'Subs.' abbreviates 'Substitution':

1. $\vdash_{\Lambda_S}$	$\varphi \rightarrow [\alpha]\langle\alpha\rangle\varphi$	C.p. of (T) for $[\alpha]$, (5) for $[\alpha]$, prop. logic
2. $\vdash_{\Lambda_S}$	$\Diamond K_\alpha\varphi \rightarrow \Diamond K_\alpha[\alpha]\langle\alpha\rangle\varphi$	1, Nec. & Subs. of (K) for K_α, modal logic
3. $\vdash_{\Lambda_S}$	$\Diamond K_\alpha[\alpha]\langle\alpha\rangle\varphi \rightarrow K_\alpha\Diamond[\alpha]\langle\alpha\rangle\varphi$	Λ_S-theorem a
4. $\vdash_{\Lambda_S}$	$K_\alpha\Diamond[\alpha]\langle\alpha\rangle\varphi \rightarrow K_\alpha\Diamond\langle\alpha\rangle\varphi$	Subs. of (T) for $[\alpha]$, modal logic, Nec. & Subs. of (K) for K_α
5. $\vdash_{\Lambda_S}$	$K_\alpha\Diamond\langle\alpha\rangle\varphi \rightarrow K_\alpha\Diamond\Diamond\varphi$	C.p. of (SET), modal logic, Nec. & Subs. of (K) for K_α
6. $\vdash_{\Lambda_S}$	$K_\alpha\Diamond\Diamond\varphi \rightarrow K_\alpha\Diamond\varphi$	C.p. of (4) for $\Box$, Nec. & Subs. of (K) for K_α
7. $\vdash_{\Lambda_S}$	$\Diamond K_\alpha\varphi \rightarrow K_\alpha\Diamond\varphi$	2, 3, 4, 5, 6, prop. logic.

(b) $\Box K_\alpha\varphi \leftrightarrow K_\alpha\Box\varphi$. Formula $(\star\star)$ $K_\alpha\Box\varphi \rightarrow \Box K_\alpha\varphi$ is obtained according to the following derivation:

1. $\vdash_{\Lambda_S}$	$K_\alpha\Box\varphi \rightarrow \Diamond K_\alpha\Box\varphi$	Subs. of (T) for $\Box$
2. $\vdash_{\Lambda_S}$	$\Diamond K_\alpha\Box\varphi \rightarrow K_\alpha\Diamond\Box\varphi$	Subs. of $(Unif - H)$
3. $\vdash_{\Lambda_S}$	$\Diamond\Box\varphi \rightarrow \Box\varphi$	C.p. of (5) for $\Box$
4. $\vdash_{\Lambda_S}$	$K_\alpha\Diamond\Box\varphi \rightarrow K_\alpha\Box\varphi$	3, Nec. & Subs. of (K) for K_α
5. $\vdash_{\Lambda_S}$	$K_\alpha\Box\varphi \rightarrow K_\alpha\Box\varphi$	1, 2, 4, prop. logic.

Formula $(\star\star\star)$ $\Box K_\alpha\varphi \rightarrow K_\alpha\Box\varphi$ is obtained according to the following derivation:

1. $\vdash_{\Lambda_S}$	$\Diamond \langle K_\alpha \rangle \Box K_\alpha\varphi \rightarrow \langle K_\alpha \rangle \Diamond\Box K_\alpha\varphi$	Subs. of c.p. of $(\star\star)$
2. $\vdash_{\Lambda_S}$	$\langle K_\alpha \rangle \Diamond\Box K_\alpha\varphi \rightarrow \langle K_\alpha \rangle K_\alpha\varphi$	Subs. of c.p. of (B) for $\Box$ modal logic
3. $\vdash_{\Lambda_S}$	$\langle K_\alpha \rangle \Diamond\Box K_\alpha\varphi \rightarrow \varphi$	2, Subs. of c.p. of (B) for K_α, prop logic
4. $\vdash_{\Lambda_S}$	$\Box \langle K_\alpha \rangle \Diamond\Box K_\alpha\varphi \rightarrow \Box\varphi$	3, Nec. & Subs. of (K) for $\Box$
5. $\vdash_{\Lambda_S}$	$\Box\Diamond \langle K_\alpha \rangle \Box K_\alpha\varphi \rightarrow \Box \langle K_\alpha \rangle \Diamond\Box K_\alpha\varphi$	1, Nec. & Subs. of (K) for $\Box$
6. $\vdash_{\Lambda_S}$	$\Diamond \langle K_\alpha \rangle \Box K_\alpha\varphi \rightarrow \Box\Diamond \langle K_\alpha \rangle \Box K_\alpha\varphi$	Subs. of (5) for $\Box$
7. $\vdash_{\Lambda_S}$	$\langle K_\alpha \rangle \Box K_\alpha\varphi \rightarrow \Diamond \langle K_\alpha \rangle \Box K_\alpha\varphi$	Subs. of (T) for $\Box$
8. $\vdash_{\Lambda_S}$	$\langle K_\alpha \rangle \Box K_\alpha\varphi \rightarrow \Box\varphi$	7, 6, 5, 4, prop. logic
9. $\vdash_{\Lambda_S}$	$K_\alpha \langle K_\alpha \rangle \Box K_\alpha\varphi \rightarrow K_\alpha\Box\varphi$	8, Nec. & Subs. of (K) for K_α
10. $\vdash_{\Lambda_S}$	$\Box K_\alpha\varphi \rightarrow K_\alpha \langle K_\alpha \rangle \Box K_\alpha\varphi$	Subs. of (B) for K_α
11. $\vdash_{\Lambda_S}$	$\Box K_\alpha\varphi \rightarrow K_\alpha\Box\varphi$	10, 9, prop. logic.

Now, substituting either $(\star\star)$ or $(\star\star\star)$ for $(Unif - H)$ in Λ_S yields the same theory. The reason is that the above derivation guarantees that $(\star\star\star)$ can be derived using $(\star\star)$, and it is the case that $(Unif - H)$ can be derived using $(\star\star\star)$, as shown by the following derivation:

1. $\vdash_{\Lambda_S}$	$\Diamond K_\alpha\varphi \rightarrow K_\alpha \langle K_\alpha \rangle \Diamond K_\alpha\varphi$	Subs. of (B) for K_α
2. $\vdash_{\Lambda_S}$	$\langle K_\alpha \rangle \Diamond K_\alpha\varphi \rightarrow \Diamond \langle K_\alpha \rangle K_\alpha\varphi$	Subs. of c.p. of $(\star\star\star)$
3. $\vdash_{\Lambda_S}$	$K_\alpha \langle K_\alpha \rangle \Diamond K_\alpha\varphi \rightarrow K_\alpha\Diamond \langle K_\alpha \rangle K_\alpha\varphi$	2, Nec. & Subs. of (K) for K_α
4. $\vdash_{\Lambda_S}$	$\langle K_\alpha \rangle K_\alpha\varphi \rightarrow \varphi$	C.p. of (B) for K_α
5. $\vdash_{\Lambda_S}$	$K_\alpha\Diamond \langle K_\alpha \rangle K_\alpha\varphi \rightarrow K_\alpha\Diamond\varphi$	4, modal logic
6. $\vdash_{\Lambda_S}$	$\Diamond K_\alpha\varphi \rightarrow K_\alpha\Diamond\varphi$	1, 3, 5, prop. logic.

(c) $\neg \odot^S_\alpha \varphi \to K_\alpha \Box \neg \odot^S_\alpha \varphi$. A derivation is as follows:

1. $\vdash_{\Lambda_S}$	$\odot^S_\alpha \varphi \to K_\alpha \Box \odot^S_\alpha \varphi$	$(s.Cl)$
2. $\vdash_{\Lambda_S}$	$\Diamond \odot^S_\alpha \varphi \to \Diamond K_\alpha \Box \odot^S_\alpha \varphi$	1, modal logic
3. $\vdash_{\Lambda_S}$	$\Diamond K_\alpha \Box \odot^S_\alpha \varphi \to K_\alpha \Diamond \Box \odot^S_\alpha \varphi$	Subs. of $(Unif - H)$
4. $\vdash_{\Lambda_S}$	$\Diamond \Box \odot^S_\alpha \varphi \to \odot^S_\alpha \varphi$	Subs. of c.p. of (B) for $\Box$
5. $\vdash_{\Lambda_S}$	$K_\alpha \Diamond \Box \odot^S_\alpha \varphi \to K_\alpha \odot^S_\alpha \varphi$	4, Nec. & Subs. of (K) for K_α
6. $\vdash_{\Lambda_S}$	$\Diamond \odot^S_\alpha \to K_\alpha \odot^S_\alpha \varphi$	2, 3, 5, prop. logic
7. $\vdash_{\Lambda_S}$	$\langle K_\alpha \rangle \Diamond \odot^S_\alpha \to \langle K_\alpha \rangle K_\alpha \odot^S_\alpha \varphi$	6, modal logic
8. $\vdash_{\Lambda_S}$	$\langle K_\alpha \rangle K_\alpha \odot^S_\alpha \varphi \to \odot^S_\alpha \varphi$	Subs. of c.p. of (B) for K_α
9. $\vdash_{\Lambda_S}$	$\neg \odot^S_\alpha \varphi \wedge \langle K_\alpha \rangle \Diamond \odot^S_\alpha \varphi \to$ $\neg \odot^S_\alpha \varphi \wedge \odot^S_\alpha \varphi$	 7, 8, prop. logic
10. $\vdash_{\Lambda_S}$	$\neg \odot^S_\alpha \varphi \wedge \langle K_\alpha \rangle \Diamond \odot^S_\alpha \varphi \to \bot$	9, prop. logic
11. $\vdash_{\Lambda_S}$	$\neg \odot^S_\alpha \varphi \to K_\alpha \Box \neg \odot^S_\alpha \varphi$	10, prop. & modal logic.

A fair question to ask at this point is why I have not presented a sound and complete proof system for *EAUST*, the full logic of objective and subjective ought-to-do's (with language $\mathcal{L}_{\mathsf{KO}}$ and semantics on *eaubt*-models). Well, there *is* a sound and complete proof system *for a variant* of this full logic, that I thoroughly explored it in my joint work with Jan Broersen (Abarca & Broersen, 2019). To be precise, in such a work a proof system Λ_{OS} is defined as follows:

- *(Axioms)* All classical tautologies from propositional logic; the **S5** schemata for $\Box$, $[\alpha]$, K_α; and schemata (SET), (IA) $(A1)$–$(A6)$, (OAC), $(Unif - H)$, (SuN), $(s.Oic)$, and $(s.Cl)$.

- *(Rules of inference) Modus Ponens*, Substitution, and Necessitation for all modal operators.

Λ_{OS} is then shown to be sound and complete with respect to a class of models known as *bi-valued eaubt*-models, which is a bigger class than the one introduced in Definition 4.18. Instead of only one function **Value**, these models include two: $\mathbf{Value}_O$, underlying the semantics for objective ought-to-do's; and $\mathbf{Value}_S$, underlying the semantics for subjective ones. Although having two value functions instead of one is extremely useful for the proof of completeness, neither the conceptual reach of this extension nor its philosophical implications have been a subject of my investigation as of yet.[27]

[27] There are some reasons to entertain skepticism about this technical extension. For instance, one could raise the following point of criticism: in bi-valued *eaubt*-models, a single history can have

Abarca and Broersen's (2019) system Λ_{OS} is sound and complete with respect to the class of bi-valued *eaubt*-models, but it is not complete with respect to *eaubt*-models. Let me briefly elaborate on this matter. As shown in Observation C.39, formulas $\odot_\alpha \varphi \to \neg \odot^S_\alpha \neg\varphi$ and $\odot^S_\alpha \varphi \to \neg \odot_\alpha \neg\varphi$, that refer to the consistency between objective and subjective ought-to-do's, are valid with respect to *eaubt*-models. It is easy to see, nonetheless, that these formulas are not valid on bi-valued *eaubt*-models, because in these models a single history can have different utilities. Thus, Λ is not complete with respect to the class of *eaubt*-models. Now, suppose that a new system Λ'_{OS} is obtained from Λ_{OS} by adding $\odot_\alpha \varphi \to \neg \odot^S_\alpha \neg\varphi$ and $\odot^S_\alpha \varphi \to \neg \odot_\alpha \neg\varphi$ as schemata. Then it is clear that Λ'_{OS} is sound with respect to *eaubt*-models, but to determine whether it is also complete is still an open problem.

4.6 Conclusion

This chapter dealt with important questions in the modelling of agency, knowledge, and obligation, on the road to building a formal theory of responsibility. I want to conclude it with a discussion of two topics: (a) an epistemic act-utilitarian stit-theoretic framework in which belief and belief-inspired obligations are accounted for, and (b) a possible extension of *EAUST* with group agency, objective group obligations, and subjective group obligations.

4.6.1 Doxastic Obligations

Recall that my study of ought-to-do is based on the premise that an agent can be excused for not meeting an obligation if it lacked knowledge that is necessary for doing so. In turn, an agent's beliefs, and the implications that these beliefs have in what the agent thinks that it should do, also amount to reasons for being excused (in those cases where the agent did not meet an obligation and faces a potential punishment). Thus, incorporating a notion of belief into *EAUST* leads to a more nuanced theory of ought-to-do, which ultimately proves useful in the construction of a rich formalization of responsibility (see Subsection 6.5.1 of Chapter 6's conclusion).

Following Bartha 2014's decision-theoretic proposal, a good way of addressing the interplay between agency, belief, and obligation—in the context of responsibility attribution—is by introducing a probabilistic semantics of belief to stit theory (see Subsection 3.5.1 of Chapter 3's conclusion). Why? Because a probabilistic

different, non-related utilities, according to either $\mathbf{Value}_O$ or $\mathbf{Value}_S$. Thus, the notion of deontic utility becomes rather vague. Observe that Definition 4.18's *eaubt*-models are particular instances of bi-valued models, where both value functions assign the same value to each history.

semantics of belief allows us to base a new sense of ought-to-do on a key concept from decision theory: expected-utility maximization. Let me refer to the obligations that arise from this new sense as *doxastic ought-to-do*'s. As presented in my joint work with Jan Broersen (Abarca & Broersen, 2021a), one can characterize these doxastic ought-to-do's as follows: agent α doxastically ought to have brought about φ iff φ is an effect of the choices that maximize α's expected deontic utility (i.e., α's rational deontically best responses).[28] To be more precise, consider the definitions below.

Definition 4.32 (Expected deontic utility). *Let $\mathcal{M}$ be a finite* eaubt-*model. For $\alpha \in Ags$, $m \in M$, and $h \in H_m$, let $\pi_\alpha[\langle m,h\rangle]$ denote α's (*ex interim*) information set at $\langle m,h\rangle$, and let μ_α be a classical discrete probability function such that, for each index $\langle m,h\rangle$, $\mu_\alpha(\pi_\alpha[\langle m,h\rangle]) > 0$ (see the discussion of μ_α in Subsection 3.5.1 of Chapter 3's conclusion, p. 112). For $L \in$* $\mathbf{Choice}_\alpha^m$*, α's* expected deontic utility of L at $\langle m,h\rangle$*, denoted by $EU_\alpha^{\langle m,h\rangle}(L)$, is defined as the value given by the following formula:*

$$EU_\alpha^{\langle m,h\rangle}(L) := \sum_{m' \sim_\alpha m, h' \in [L]_\alpha^{m'}} \mu_\alpha\left(\{\langle m',h'\rangle\} \mid \pi_\alpha\left[\langle m',h'\rangle\right]\right) \cdot \mathbf{Value}(h'),$$

where recall that I write $m \sim_\alpha m'$ if there exist $h \in H_m$ and $h' \in H_{m'}$ such that $\langle m,h\rangle \sim_\alpha \langle m',h'\rangle$.

This means that α's expected deontic utility of an available choice L at $\langle m,h\rangle$ is calculated by summing the utilities of all the histories lying in the epistemic clusters of L, weighted by the probabilities that α assigns to the indices anchored by those histories, conditional on α's information set at those indices.[29] Observe, then, that, for all $\alpha \in Ags$, $m \in M$, and $L \in \mathbf{Choice}_\alpha^m$, $EU_\alpha^{\langle m,h\rangle}(L) = EU_\alpha^{\langle m,h'\rangle}(L)$ for every $h, h' \in H_m$.

Definition 4.33 (Maximal expected deontic utility). *Let $\mathcal{M}$ be a finite* eaubt-*model with probability functions μ_α (where α ranges over Ags), defined just as in Definition 4.32. Since M is finite, then for all $\alpha \in Ags$, $m \in M$, and $h \in H_m$, the set $\left\{EU_\alpha^{\langle m,h\rangle}(L); L \in \mathbf{Choice}_\alpha^m\right\}$ has a maximum. Therefore, at $\langle m,h\rangle$ there are actions that maximize α's expected deontic utility, namely the ones whose expected deontic utility is the same as said maximum. The set of such actions is denoted by* $\mathbf{EU}_\alpha^{\langle m,h\rangle}$.

[28] For a logic-based study of agents' beliefs and their rationality (as ensuing from best responses in games), the reader is referred to Bjorndahl, Halpern, and Pass (2011, 2017).

[29] The reason for conditioning on information sets of the form $\pi[\langle m',h'\rangle]$ in Definition 4.32 lies in the kind of belief—in the context of the stages of information disclosure—that Abarca and Broersen (2021a) wanted to base expectation on. Recall that *ex interim* knowledge is presently captured by $K_\alpha\varphi$. Since Abarca and Broersen's p-1 belief refines *ex interim* knowledge and thus should be seen as *ex interim* belief, expected deontic utility is based on *degrees* of *ex interim* belief, so that conditioning with respect to $\pi[\langle m',h'\rangle]$ is justified.

With the idea of maximal expected deontic utility, doxastic ought-to-do's can be formally introduced. Let the modality $\odot^{\mathcal{B}}_{\alpha}\varphi$ express that agent α doxastically ought to have seen to it that φ. Then $\mathcal{M}, \langle m, h\rangle \models \odot^{\mathcal{B}}_{\alpha}\varphi$ iff for all $L \in \mathbf{EU}^{\langle m,h\rangle}_{\alpha}$, $[L]^{m'}_{\alpha} \subseteq \left|\varphi\right|^{m'}$ for every m' such that $m \sim_{\alpha} m'$. In other words, at an index an agent doxastically ought to have seen to it that φ iff φ is an effect of all the epistemic clusters of those actions that maximized the agent's expected deontic utility at said index.[30] Perhaps the best way of exploring this doxastic sense of obligation is by means of an example.

Recall Example 3.12 (p. 112), the anesthesiologist example in Subsection 3.5.1 of Chapter 3's conclusion. This example involves a doctor who supplied anesthetics to a patient before a surgery. The patient had eaten just before the surgery, and the doctor did not know this. Anesthetics must have been supplied only on an empty stomach, so the patient died due to the interaction between the food and the anesthetics. A small variation of the example, catered to the needs of illustrating doxastic ought-to-do's, is included in Figure 4.9.

As implied by the statement of the example, h_3 is the actual history. Suppose that *doctor*'s doxastic state is now given by the discrete probability function $\mu_{doctor} : 2^{I(M\times H)} \to [0,1]$, where $\mu_{doctor}\left(\left\langle m_i, h_j\right\rangle\right) = \frac{0.9}{4}$ for all $i, j \in \{1,2\}$, $\mu_{doctor}\left(\left\langle m_i, h_j\right\rangle\right) = \frac{0.1}{4}$ for all $i \in \{1,3\}$ and $j \in \{3,4\}$. In the diagram, this is represented by labelling the left-hand side of the model with the tag $\mu_{doctor} : 0.9$, and the right-hand side with the tag $\mu_{doctor} : 0.1$. Thus, $\mathcal{M}, \langle m_3, h_3\rangle \models \neg B_{doctor}\neg e$: *at the actual index* doctor *did not p-1 believe that the patient had not eaten*. This is due to the following arguments. *Doctor*'s information set at the actual index is $\pi_{doctor}\left[\langle m_3, h_3\rangle\right] = \{\langle m_3, h_3\rangle, \langle m_2, h_1\rangle\}$. In such an information set, the set of indices at which the patient has not eaten is $||\neg e|| \cap \pi_{doctor}\left[\langle m_3, h_3\rangle\right] = \{\langle m_2, h_1\rangle\}$; at the actual index, the probability that *doctor*

[30] Interestingly, Horty (2001, Chapter 4) had already considered the possibility of formalizing obligation using an agent's expected value of its actions. Horty's brief assessment of this possibility presupposed that probability distributions would be given over the set of histories in *eaubt*-frames. The probability assigned to a history within an action cell would represent "its chance of occurring should the agent choose to perform the action." The expected value of each available action—obtained by summing the values of the histories of the action, each weighted by the probability assigned to its history—would then provide a preference ordering on the set of actions, and an agent would "ought to see to it that some proposition holds whenever doing so is a necessary condition for performing any action whose expected value is among the greatest available" (Horty, 2001, Chapter 4, p. 59). Horty disfavored the approach because, in his opinion, it relied on probabilistic information about the outcomes that is often unavailable or meaningless. However, if one interprets the probability distributions as representations of subjective belief, then the resulting framework *does not* rely on information that is often unavailable or meaningless. Rather, it incorporates the well-known theory of probability-based belief into act-utilitarian stit theory, leading to a different, doxastic sense of ought.

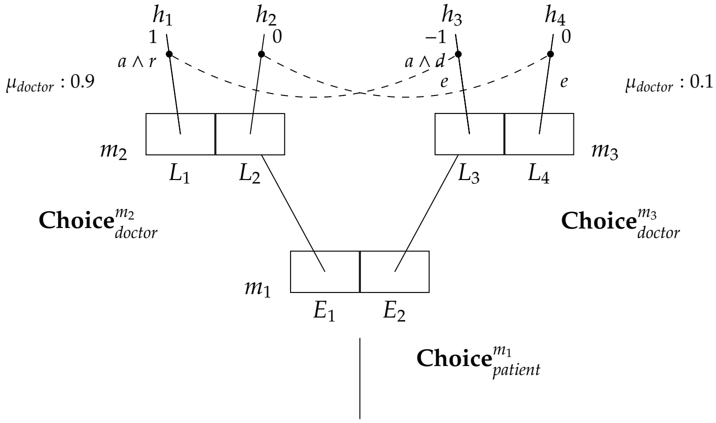


Figure 4.9: *Anesthesiologist example, once more.*

assigned to the event that the patient had not eaten, conditional on her information set, is $\mu_{doctor}\,(||\neg e||\mid \pi_{doctor}\,[\langle m_3,h_3\rangle]) = \frac{\mu_{doctor}(||\neg e|| \cap \pi_{doctor}[\langle m_3,h_3\rangle])}{\mu_{doctor}(\pi_{doctor}[\langle m_3,h_3\rangle])} = \frac{\frac{0.9}{4}}{\frac{0.9+0.1}{4}} = 0.9 \neq 1$. Thus, *doctor* did not p-1 believe that the patient had not eaten.

Now, let us calculate *doctor*'s expected deontic utility for her available choices:

- As for the choices available at m_2, observe that, for all $i \in \{1,2\}$,

$$\begin{aligned} EU^{\langle m_2,h_i\rangle}_{doctor}(L_1) &= \mu_{doctor}\,(\langle m_2,h_1\rangle \mid \pi_{doctor}\,[\langle m_2,h_1\rangle]) \cdot \mathbf{Value}(h_1) + \\ &\quad \mu_{doctor}\,(\langle m_3,h_3\rangle \mid \pi_{doctor}\,[\langle m_3,h_3\rangle]) \cdot \mathbf{Value}(h_3) \\ &= 0.9 \cdot 1 + 0.1 \cdot (-1) = 0.8. \end{aligned}$$

$$\begin{aligned} EU^{\langle m_2,h_i\rangle}_{doctor}(L_2) &= \mu_{doctor}\,(\langle m_2,h_2\rangle \mid \pi_{doctor}\,[\langle m_2,h_2\rangle]) \cdot \mathbf{Value}(h_2) + \\ &\quad \mu_{doctor}\,(\langle m_3,h_4\rangle \mid \pi_{doctor}\,[\langle m_3,h_4\rangle]) \cdot \mathbf{Value}(h_4) \\ &= 0.9 \cdot 0 + 0.1 \cdot 0 = 0. \end{aligned}$$

- As for the choices available at m_3, observe that, for all $i \in \{3,4\}$, $EU^{\langle m_3,h_i\rangle}_{\alpha}(L_3) = EU^{\langle m_2,h_1\rangle}_{\alpha}(L_1)$ and $EU^{\langle m_3,h_i\rangle}_{\alpha}(L_4) = EU^{\langle m_2,h_2\rangle}_{\alpha}(L_2)$.

Therefore, $\mathbf{EU}^{\langle m_2,h_1\rangle}_{\alpha} = \mathbf{EU}^{\langle m_2,h_2\rangle}_{\alpha} = \{L_1\}$, and $\mathbf{EU}^{\langle m_3,h_3\rangle}_{\alpha} = \mathbf{EU}^{\langle m_3,h_4\rangle}_{\alpha} = \{L_3\}$. This implies that, for all $i \in \{2,3\}$ and $h \in H_{m_i}$, $\mathcal{M}, \langle m_i, h\rangle \models \odot^{\mathcal{B}}_{doctor} a$: *at all indices based on m_2 and m_3* doctor *doxastically ought to have supplied the anesthetics*. Furthermore, observe that, for all $i \in \{2,3\}$ and $h \in H_{m_i}$, $\mathcal{M}, \langle m_i, h\rangle \models K_\alpha \odot^{\mathcal{B}}_{doctor} a$: *at all indices based on m_2 and m_3* doctor *knew that she doxastically ought to have supplied the anesthetics*. Thus, even if *doctor* did not have p-1 certainty that the patient had not eaten, she still knew that she doxastically ought to have supplied the anesthetics. Coupled with the facts that *doctor* did not know that the patient had eaten and that *doctor* did not knowingly kill the patient, the satisfaction of these last two formulas at the actual index provides a good reason for excusing *doctor* from having moral responsibility of the patient's death, despite having caused it.[31]

To explore the link between probability-based belief and obligation is a natural step to take in the line both of Horty's (2001) *AUST* and of the extensions with epistemic modalities that were presented in this chapter. To clarify, recall that Horty's act-utilitarian ought-to-do is based on a measure of dominance for choices of action. Extending the theory of ought-to-do with p-1 belief and with the different notion of optimality that stems from expected utility, then, adds a new dimension to the discussion. A very interesting problem for future work along these lines, then, concerns implementing the ideas of *belief revision*—in terms of conditional belief—to formalize conditional doxastic ought-to-do's. The intuition is that if at an index an agent has learned that ψ is the case then the doxastic obligations that such an agent has at the index should in principle be subject to the revision with ψ—just as beliefs are. Formulas of the form $\odot_\alpha^{\mathcal{B}/\psi}\varphi$ could then capture these revised doxastic ought-to-do's, and possible semantics for these formulas could depend on the restriction of the model's indices to those where ψ holds (just as happens for the version of conditional belief discussed by Abarca and Broersen (2021a)). In fact, for good pointers in this respect the reader is referred to Horty (2001, Chapter 4), where a stit-theoretic account of conditional ought-to-do's is presented. Of course, this is all the more interesting and relevant in the context of building a finespun theory of responsibility.

4.6.2 Group Obligations

Collective responsibility refers to a relation between a group of agents and some state of affairs such that the group is responsible for the state of affairs iff the group's degree of involvement in the realization of that state of affairs warrants

[31] As for the logic-based properties of doxastic ought-to-do's, one can adapt the arguments of Abarca and Broersen (2021a) to show that $\odot_\alpha^{\mathcal{B}}$ is a **KD45** operator for which the validity of the following formulas additionally holds: $\odot_\alpha^{\mathcal{B}}\varphi \to \odot_\alpha^{\mathcal{B}} (K_\alpha[\alpha]\varphi)$, $\odot_\alpha^{\mathcal{B}}\varphi \to \Diamond K_\alpha\varphi$ (a doxastic version of Kant's directive of *ought implies can*), $\odot_\alpha^{\mathcal{B}}\varphi \to K_\alpha\Box \odot_\alpha^{\mathcal{B}} \varphi$, and $\neg \odot_\alpha^{\mathcal{B}} \varphi \to K_\alpha\Box\neg \odot_\alpha^{\mathcal{B}} \varphi$. Interestingly, if one builds a logic including all three ought-to-do modalites (objective, subjective, and doxastic), then doxastic and objective ought-to-do's are not necessarily consistent ($\not\models \odot_\alpha^{\mathcal{B}}\varphi \to \neg \odot_\alpha \neg\varphi$ and $\not\models \odot_\alpha\varphi \to \neg \odot_\alpha^{\mathcal{B}} \neg\varphi$), and doxastic and subjective ought-to-do's neither ($\not\models \odot_\alpha^{\mathcal{S}}\varphi \to \neg \odot_\alpha^{\mathcal{B}} \neg\varphi$ and $\not\models \odot_\alpha^{\mathcal{B}}\varphi \to \neg \odot_\alpha^{\mathcal{S}} \neg\varphi$). This highlights the discrepancy between dominance and expected-utility maximization, just as evidenced by the famous thought experiment in decision theory known as *Newcomb's problem* (Ahmed, 2018; Eells, 1982; Gibbard & Harper, 1978; Nozick, 1969; Weirich, 2020). As for metalogic results, one can once again adapt the proofs of Abarca and Broersen (2021a) to show that there is a sound and complete proof system for the full logic of objective, subjective, and doxastic ought-to-do's.

collective blame or collective praise. An important discussion in responsibility attribution concerns how individual and collective responsibility relate to one another. In the exploration of such a relation, the concept of *group obligations* is key (Duijf, 2018, Chapters 2 & 3).

Recall from Chapter 2's Subsection 2.4.1 that *atemporal group stit theory* is the extension of atemporal *BST* with modalities of the form $[G]\varphi$ (where $G \subseteq Ags$), meant to express that coalition G has seen to it that φ. As mentioned there, the semantics for $[G]\varphi$ on *bt*-models depends on *joint actions*. For *bt*-model $\mathcal{M}$, coalition $G \subseteq Ags$, and $m \in M$, G's set of available joint actions at m is defined as $\mathbf{Choice}^m_G := \{\bigcap_{\alpha \in G} \mathbf{Choice}^m_\alpha(h); h \in H_m\}$, so that $\mathcal{M}, \langle m, h \rangle \models [G]\varphi$ iff $\mathcal{M}, \langle m, h' \rangle \models \varphi$ for every $h' \in \mathbf{Choice}^m_G(h)$. One can then think of an extension of atemporal *BST* with objective group obligations, underlying a modality of the form $\odot_G\varphi$. Following Horty's (2001, Chapter 6) ideas, the semantics for $\odot_G\varphi$ can be based on dominance rankings over joint actions:

Definition 4.34 (Objective dominance ordering over joint choices). *For an* eaubt-*frame with M as its set of moments, $G \subseteq Ags$, and $m \in M$, consider the following definitions:*

- *Let* $\mathbf{State}^m_G := \left\{S \subseteq H_m; S = \bigcap_{\beta \in Ags - G} s(\beta), \textit{ for } s \in \mathbf{Select}^m\right\}$.
- *Let $\leq$ be an ordering on $\mathbf{Choice}^m_G$ defined by the rule: for $L, L' \in \mathbf{Choice}^m_G$, $L \leq L'$ iff for all $S \in \mathbf{State}^m_G$, $L \cap S \leq L' \cap S$, where recall that $\leq$ is an ordering on 2^{H_m} such that $X \leq Y$ iff $\mathbf{Value}(h) \leq \mathbf{Value}(h')$ for every $h \in X$ and $h' \in Y$. I write $L \prec L'$ iff $L \leq L'$ and $L' \nleq L$.*
- *Let* $\mathbf{Optimal}^m_G := \left\{L \in \mathbf{Choice}^m_G; \text{there is no } L' \in \mathbf{Choice}^{m_*}_G \text{such that } L \prec L'\right\}$.

For an *eaubt*-model $\mathcal{M}$, then, one can set that $\mathcal{M}, \langle m, h \rangle \models \odot_G\varphi$ iff for all $L \in \mathbf{Choice}^m_G$ such that $\mathcal{M}, \langle m, h_L \rangle \not\models \varphi$ for some $h_L \in L$, there is $L' \in \mathbf{Choice}^m_G$ such that $L \prec L'$ and, if $L'' = L$ or $L' \leq L''$, then $\mathcal{M}, \langle m, h' \rangle \models \varphi$ for every $h' \in L''$.[32] Furthermore, the strategy of focusing on dominance of joint actions can also help us formalize *subjective group obligations*—according to the intuitions about subjective ought-to-do's advanced in this chapter. In this case, however, one must decide what kind of group knowledge is to be used. Inspired by the literature on epistemic logic and by the idea that group agency is a distributed process, here I mention one candidate: *distributed knowledge*, expressed by modality $D_G\varphi$ (for $G \subseteq Ags$) (see, for instance, Fagin et al., 1995; Gerbrandy, 1998; Halpern & Fagin,

[32] Similar to what happens for individual obligations, for finite-choice *eaubt*-models this clause is equivalent to $\mathcal{M}, \langle m, h \rangle \models \odot_\alpha\varphi$ iff for all $L \in \mathbf{Optimal}^m_G$, $\mathcal{M}, \langle m, h' \rangle \models \varphi$ for every $h' \in L$. Furthermore, both semantics imply that $\odot_\alpha\varphi \leftrightarrow \odot_{\{\alpha\}}\varphi$ is valid for every $\alpha \in Ags$.

1989, for a thorough examination of distributed knowledge).[33] Similar to what happens for individual agents, this kind of group knowledge leads to the idea of a joint action's distributed epistemic equivalents:

Definition 4.35 (Distributed epistemic clusters for coalitions). *Let $\mathcal{M}$ be an* eaubt-*frame with M as its set of moments. Let $G \subseteq Ags$, and let $m, m' \in M$ be such that $m\,(\bigcap_{\alpha\in G} \sim_\alpha)\,m'$. For $L \in \mathbf{Choice}^m_G$, L's* distributed epistemic cluster at m' *is the set*

$$[L]^{m'}_G := \left\{ h' \in H_{m'};\ \textit{there is}\ h \in L\ \textit{s. t.}\ \ \langle m,h\rangle \left(\bigcap_{\alpha\in G} \sim_\alpha \right) \langle m',h'\rangle \right\}.$$

Thus, subjective dominance orderings can be defined over joint choices analogously to how they were defined for individual agents in Section 4.4:

Definition 4.36 (Subjective dominance ordering over joint choices). *For an* eaubt-*frame with M as its set of moments, $G \subseteq Ags$, and $m \in M$, consider the following definitions:*

- *Let $\preceq_s$ be an ordering on $\mathbf{Choice}^m_G$ defined by the rule: for $L, L' \in \mathbf{Choice}^m_G$, $L \preceq_s L'$ iff for all m' such that $m\,(\bigcap_{\alpha\in G} \sim_\alpha)\,m'$ and each $S \in \mathbf{State}^{m'}_G$, $[L]^{m'}_G \cap S \leq [L']^{m'}_G \cap S$. I write $L \prec_s L'$ iff $L \preceq_s L'$ and $L' \npreceq_s L$.*
- *Let* $\mathbf{SOptimal}^m_G := \left\{ L \in \mathbf{Choice}^m_G;\ \text{there is no } L' \in \mathbf{Choice}^m_G \text{ s. t. } L \prec_s L' \right\}$.

For an *eaubt*-model $\mathcal{M}$, one can then set that $\mathcal{M}, \langle m,h\rangle \models \odot^S_G \varphi$ iff for all $L \in \mathbf{Choice}^m_G$ such that $\mathcal{M}, \langle m', h_L\rangle \not\models \varphi$ for some m' such that $m\,(\bigcap_{\alpha\in G} \sim_\alpha)\,m'$ and some $h_L \in [L]^{m'}_G$, there is $L' \in \mathbf{Choice}^m_G$ such that $L \prec_s L'$ and, if $L'' = L'$ or $L' \preceq_s L''$, then $\mathcal{M}, \langle m'', h''\rangle \models \varphi$ for every m'' such that $m\,(\bigcap_{\alpha\in G} \sim_\alpha)\,m''$ and every $h'' \in [L'']^{m''}_G$.[34]

To illustrate both senses of group ought-to-do's, consider Chapter 3's Example 3.4 (p. 81), where a bomb squad is trying to defuse a bomb. Recall that there were two cases: case a was depicted in Figure 3.2 (p. 86), and case b was depicted in Figure 3.3 (p. 87). To turn both these figures' models into *eaubt*-models (with group notions), let **Value** be a function defined on their sets of histories such that $\mathbf{Value}(h) = 1$ on all histories where the bomb squad has defused the bomb (i.e. $h \in \{h_2, h_7, h_9\}$), and $\mathbf{Value}(h) = 0$ on all other histories. Let $G = \{Luther, Benji\}$. As implied by the statement of the example, the

[33] It would also be interesting to explore subjective group ought-to-do's based on common knowledge.

[34] Just as in the case of objective group ought-to-do's, in finite-choice *eaubt*-models such a truth condition is equivalent to $\mathcal{M}, \langle m,h\rangle \models \odot^S_\alpha \varphi$ iff for all $L \in \mathbf{SOptimal}^m_G$, $\mathcal{M}, \langle m',h'\rangle \models \varphi$ for every m' such that $m\,(\bigcap_{\alpha\in G} \sim_\alpha)\,m'$ and every $h' \in [L]^{m''}_\alpha$. Furthermore, both semantics imply that $\odot^S_\alpha \varphi \leftrightarrow \odot^S_{\{\alpha\}} \varphi$ is valid for every $\alpha \in Ags$.

actual history is h_{10}. Observe, then, that $\mathbf{Optimal}_G^{m_4} = \{G_L \cap R_B\}$ in both case a and case b, so that the objectively optimal joint action for group G at m_4 is given by *Luther*'s cutting the green wire of his bomb and *Benji*'s cutting the red wire of his. As such, if s stands for the proposition 'the bombs are defused,' and using Chapter 3's 'next' operator X to remain consistent with that chapter's presentation, then $\mathcal{M}, \langle m_4, h_{10}\rangle \models \odot_G Xs$: *at* $\langle m_4, h_{10}\rangle$ *the group made up of* Luther *and* Benji *objectively ought to have defused the bomb.* As for subjective obligations, observe that in case a $\mathbf{SOptimal}_G^{m_4} = \{R_L \cap R_B, R_L \cap G_B, G_L \cap R_B\}$, which implies that $\mathcal{M}, \langle m_4, h_{10}\rangle \models \neg \odot_G^S Xs$: *at* $\langle m_4, h_{10}\rangle$ *the group made up of* Luther *and* Benji *did not subjectively ought to have defused the bomb.* In contrast, in case b $\mathbf{SOptimal}_G^{m_4} = \{G_L \cap R_B\}$, so that $\mathcal{M}, \langle m_4, h_{10}\rangle \models \odot_G^S Xs$: *at* $\langle m_4, h_{10}\rangle$ *the group subjectively ought to have defused the bomb.*[35]

When group notions are being discussed, it is always important to reflect on the relation between the group and its members. In the case of objective and subjective group ought-to-do's, obligations are neither *downward hereditary* nor *upward hereditary* (Horty, 2001, Chapter 6, p. 131). To clarify, take $G \subseteq Ags$ and $\alpha \in G$. To see that downward inheritance fails, suppose that G is objectively, resp. subjectively, obligated to see to it that φ and that a necessary condition for G to see to it that φ is that α sees to it that ψ. In terms of formulas, $\mathcal{M}, \langle m, h\rangle \models \odot_G \varphi \wedge ([G]\varphi \rightarrow [\alpha]\psi)$, resp. $\mathcal{M}, \langle m, h\rangle \models \odot_G^S \varphi \wedge ([G]\varphi \rightarrow [\alpha]\psi)$. Then it is not necessarily true that α is objectively, resp. subjectively, obligated to see to it that ψ: $\neg \odot_\alpha \psi$, resp. $\neg \odot_\alpha^S \psi$, might hold at $\langle m, h\rangle$. Thus, $\not\models (\odot_G \varphi \wedge ([G]\varphi \rightarrow [\alpha]\psi)) \rightarrow \odot_\alpha \psi$ and $\not\models \left(\odot_G^S \varphi \wedge ([G]\varphi \rightarrow [\alpha]\psi)\right) \rightarrow \odot_\alpha^S \psi$. A counterexample for both objective and subjective group ought-to-do's is given by Horty's (2001, Chapter 6, Figure 6.1, p. 133) 'swimming pool example.' To see that upward inheritance fails, suppose that α is objectively, resp. subjectively, obligated to see to it that φ and that a necessary condition for α to see to it that φ is that G sees to it that ψ. In terms of formulas, $\mathcal{M}, \langle m, h\rangle \models \odot_\alpha \varphi \wedge ([\alpha]\varphi \rightarrow [G]\psi)$, resp. $\mathcal{M}, \langle m, h\rangle \models \odot_\alpha^S \varphi \wedge ([\alpha]\varphi \rightarrow [G]\psi)$. Then it is not necessarily the case that G is objectively, resp. subjectively, obligated to see to it that ψ: $\neg \odot_G \psi$, resp. $\neg \odot_G^S \psi$, might hold at $\langle m, h\rangle$. Thus, $\not\models (\odot_\alpha \varphi \wedge ([\alpha]\varphi \rightarrow [G]\psi)) \rightarrow \odot_G \psi$ and $\not\models \left(\odot_\alpha^S \varphi \wedge ([\alpha]\varphi \rightarrow [G]\psi)\right) \rightarrow \odot_G^S \psi$. A counterexample for both objective and subjective group ought-to-do's was also given by Horty (2001, Chapter 5, Figure 5.5, p. 116).

[35] As for the logic-based properties of group ought-to-do's, it is easy to verify that both $\odot_G$ and $\odot_G^S$ turn out to be **KD45** operators. Moreover, the following formulas turn out to be valid: $\Box\varphi \rightarrow \odot_G \varphi$, $\odot_G \varphi \rightarrow \Box \odot_G \varphi$, $\odot_G \varphi \rightarrow \odot_G ([G]\varphi)$, $\odot_G \varphi \rightarrow \Diamond[G]\varphi$, $D_G \varphi \rightarrow [G]\varphi$, $K_\beta \varphi \rightarrow D_G \varphi$ for every $\beta \in G$, $\Diamond D_G \varphi \rightarrow D_G \Diamond \varphi$, $D_G \Box \varphi \rightarrow \odot_G^S \varphi$, $\odot_G^S \varphi \rightarrow D_G \Box \odot_G^S \varphi$, $\odot_G^S \varphi \rightarrow \odot_G^S (D_G \varphi)$, and $\odot_G^S \varphi \rightarrow \Diamond D_G \varphi$. However, as far as metalogic results go, observe that, since the logic extends atemporal group stit theory, then it is not finitely axiomatizable (Herzig & Schwarzentruber, 2008) (see also Chapter 2's Subsection 2.4.1).

A possible way of mending the disparity between group and individual obligations is through Duijf's (2018, Chapter 2) *member obligations*. Intuitively, a member obligation is what a group member ought to do to help ensure that the group fulfills its group obligation. Thus, one can set the following definitions: (a) if group G is objectively obligated to see to it that φ and it is historically necessary that G will see to it that φ iff α sees to it that ψ, then α has the objective member-obligation to see to it that ψ; and (b) if group G is subjectively obligated to see to it that φ and G distributively knows that G will see to it that φ iff α sees to it that ψ, then α has the subjective member-obligation to see to it that ψ. In this way, member obligations guarantee that both downward and upward inheritance is satisfied. To incorporate member obligations into $EAUST$, new modalities would be needed. A logic-based exploration of such an extension is an interesting path for future work, specially in order to tackle fine-grained responsibility-related problems. For instance, suppose that a group failed to comply with one of their group obligations and thus is collectively blameworthy. If the group distributively knew that they would have fulfilled their group obligation iff each member had played their part (or member obligation), then an afterwards deliberation would lead to the group's identifying the members who failed to play their part, so that the group could hold them responsible to an appropriate degree at the *ex post* stage (Duijf, 2018, Chapter 3, p. 135).

Appendix B A Correspondence Result

Proposition 4.24. *Let $\mathcal{F}$ be a labelled* eaubt*-frame, and let $\mathcal{F}'$ be its transform structure. Then $\mathcal{F}'$ is an unconstrained* eaubt*-frame. Additionally, if $\mathcal{F}$ is Horty-like, then $\mathcal{F}'$ is an* eaubt*-frame, and if $\mathcal{F}$ is finite-choice, then $\mathcal{F}'$ is also finite-choice.*

Proof. From Definition 4.23, one can see that if $\mathcal{F}$ is a labelled *eaubt*-frame then $\mathcal{F}'$ is an unconstrained *eaubt*-frame. Now, assume that $\mathcal{F}$ is Horty-like. To show that $\mathcal{F}'$ is an *eaubt*-frame, one needs to show that $\mathcal{F}'$ satisfies constraints ($\mathtt{OAC}$) and ($\mathtt{Unif-H}$).

For ($\mathtt{OAC}$), let $\mathcal{M}'$ be any model based on $\mathcal{F}'$. Assume that $\mathcal{M}', \langle m,h\rangle \models K_\alpha\varphi$. For all $h' \in \mathbf{Choice}'^{m}_{\alpha}$, $Lbl_\alpha(\langle m,h\rangle) = Lbl_\alpha(\langle m,h'\rangle)$. Constraint (C4) on $\mathcal{F}$ and reflexivity of $\sim_\alpha$ imply that $\langle m,h\rangle \sim_\alpha \langle m,h'\rangle$. Therefore, the definition of $\sim'_\alpha$ implies that $\langle m,h\rangle \sim'_\alpha \langle m,h'\rangle$. The main assumption then implies that $\mathcal{M}', \langle m,h'\rangle \models \varphi$. Since $h' \in \mathbf{Choice}'^{m}_{\alpha}$, this in turn implies that $\mathcal{M}', \langle m,h\rangle \models [\alpha]\varphi$. Therefore, formula $K_\alpha\varphi \to [\alpha]\varphi$ is valid on $\mathcal{M}'$ and thus valid on $\mathcal{F}'$. According to item 5a in the list of *EAUST*'s logic-based properties (p. 172), this formula defines ($\mathtt{OAC}$), so that its validity on $\mathcal{F}'$ implies that $\mathcal{F}'$ satisfies ($\mathtt{OAC}$).

For ($\mathtt{Unif-H}$), let $\mathcal{M}'$ be any model based on $\mathcal{F}'$. Assume that $\mathcal{M}', \langle m,h\rangle \models \Diamond K_\alpha\varphi$. This means that there is $h_* \in H_m$ such that $\mathcal{M}', \langle m,h_*\rangle \models K_\alpha\varphi$. Take $\langle m',h'\rangle$ such that $\langle m,h\rangle \sim'_\alpha \langle m',h'\rangle$. The definition of $\sim'_\alpha$ implies that $\langle m,h\rangle \sim_\alpha \langle m',h'\rangle$ and that $Lbl_\alpha(\langle m,h\rangle) = Lbl_\alpha(\langle m',h'\rangle)$. Constraint (C4) on $\mathcal{F}$ and reflexivity of $\sim_\alpha$ imply that $\langle m,h_*\rangle \sim_\alpha \langle m,h\rangle$, so that transitivity of $\sim_\alpha$ implies that $\langle m,h_*\rangle \sim_\alpha \langle m',h'\rangle$. On the one hand, since $\mathcal{F}$ satisfies ($\mathtt{UAAT}$) and $\langle m,h_*\rangle \sim_\alpha \langle m',h'\rangle$, then there is $h'_* \in H_{m'}$ such that ($\star$) $Lbl_\alpha(\langle m,h_*\rangle) = Lbl_\alpha(\langle m',h'_*\rangle)$. On the other hand, constraint (C4) implies that ($\star\star$) $\langle m,h_*\rangle \sim_\alpha \langle m',h'_*\rangle$. Therefore, ($\star$) and ($\star\star$) imply that $\langle m,h_*\rangle \sim'_\alpha \langle m',h'_*\rangle$. The assumption that $\mathcal{M}', \langle m,h_*\rangle \models K_\alpha\varphi$ then implies that $\mathcal{M}', \langle m',h'_*\rangle \models \varphi$, which gives that $\mathcal{M}', \langle m',h'\rangle \models \Diamond\varphi$ and thus that $\mathcal{M}', \langle m,h\rangle \models K_\alpha\Diamond\varphi$. Therefore, formula $\Diamond K_\alpha\varphi \to K_\alpha\Diamond\varphi$ is valid on $\mathcal{M}'$ and thus valid on $\mathcal{F}'$. According to item 6a in the list of *EAUST*'s logic-based properties (p. 173), this formula defines ($\mathtt{Unif-H}$), so its validity on $\mathcal{F}'$ implies that $\mathcal{F}'$ satisfies ($\mathtt{Unif-H}$).

As for finite-choice labelled *eaubt*-frames, it is clear that if in $\mathcal{M}$ **Choice** maps each agent α and moment m to a *finite* partition $\mathbf{Choice}^{m}_{\alpha}$ of H_m then so does **Choice**$'$ in $\mathcal{F}'$.

□

Lemma B.37. *Let $\mathcal{M}$ be a Horty-like finite-choice labelled* eaubt*-model, and let $\mathcal{M}'$ be its transform finite-choice* eaubt*-model. Let us redefine $\leq_s$ in $\mathcal{M}'$ so that, for $\alpha \in Ags$, $m \in M$, and $L, L' \in \mathbf{Choice}^{m}_{\alpha}$, $L \leq_s L'$ iff for all m' such that $m \sim_\alpha m'$, $[L]^{m'}_{\alpha} \leq [L']^{m'}_{\alpha}$. Then the following points hold:*

(i) For all $\alpha \in Ags$ and $m, m' \in M$, $m \sim_\alpha m'$ iff $m \sim'_\alpha m'$.

*(ii) For all $\alpha \in Ags$, $m \in M$, and $L \in$ **Choice**$^m_\alpha$, $[L]^{m'}_\alpha = Exe^{m'}_\alpha(Lbl(L))$ for every m' such that $m \sim'_\alpha m'$.*

(iii) For all $\alpha \in Ags$, $m \in M$, and $\tau \in Tps^m_\alpha$, $[Exe^m_\alpha(\tau)]^{m'}_\alpha = Exe^{m'}_\alpha(\tau)$ for every m' such that $m \sim'_\alpha m'$.

*(iv) For all $\alpha \in Ags$, $m \in M$, and $L \in$ **Choice**$^m_\alpha$, $L \in$ **SOptimal**$^m_\alpha$ iff $Lbl(L) \in TOptimal^m_\alpha$.*

Proof. (i) To show that this point is true, recall that we write $m \sim'_\alpha m'$ if there exist $h \in H_m$ and $h' \in H_{m'}$ such that $\langle m, h\rangle \sim'_\alpha \langle m', h'\rangle$. For the left-to-right implication, observe that if $m \sim_\alpha m'$ then constraint (UAAT) of $\mathcal{M}$ ensures that there exist $h \in H_m$ and $h' \in H_{m'}$ such that $Lbl_\alpha(\langle m, h\rangle) = Lbl_\alpha(\langle m', h'\rangle)$. By definition of $\sim'_\alpha$, this implies that $\langle m, h\rangle \sim'_\alpha \langle m', h'\rangle$ and thus that $m \sim'_\alpha m'$. For the right-to-left implication, observe that the definition of $\sim'_\alpha$ and constraint (C4) straightforwardly imply that if $m \sim'_\alpha m'$ then $m \sim_\alpha m'$.

(ii) First, observe that point i above, with constraint (UAAT), ensures that $Exe^{m'}_\alpha(Lbl(L))$ is defined for every m' such that $m \sim'_\alpha m'$ (since the fact that $m \sim'_\alpha m'$ implies that $m \sim_\alpha m'$). For the $\subseteq$ inclusion, assume that $h' \in [L]^{m'}_\alpha$. This means that there exists $h \in L$ such that $\langle m, h\rangle \sim'_\alpha \langle m', h'\rangle$. This last fact implies that $Lbl_\alpha(\langle m', h'\rangle) = Lbl_\alpha(\langle m, h\rangle) = Lbl(L)$. Condition (EL) of $\mathcal{M}$ (see Definition 2.24) ensures that $h' \in Exe^{m'}_\alpha(Lbl_\alpha(\langle m', h'\rangle))$, so that the last equality shows that $h' \in Exe^{m'}_\alpha(Lbl(L))$. For the $\supseteq$ inclusion, assume that $h' \in Exe^{m'}_\alpha(Lbl(L))$. By condition (LE) of $\mathcal{M}$ (see Definition 2.24), this implies that $Lbl_\alpha(\langle m', h'\rangle) = Lbl(L)$. This last equality, coupled with the fact that $m \sim_\alpha m'$, implies that, for any $h \in L$, $\langle m, h\rangle \sim'_\alpha \langle m', h'\rangle$. Therefore, $h' \in [L]^{m'}_\alpha$.

(iii) For the $\subseteq$ inclusion, assume that $h' \in [Exe^m_\alpha(\tau)]^{m'}_\alpha$. This means that there exists $h \in Exe^m_\alpha(\tau)$ such that $\langle m, h\rangle \sim'_\alpha \langle m', h'\rangle$. This last fact implies, on the one hand, that $m \sim_\alpha m'$, so that constraint (UAAT) of $\mathcal{M}$ yields that $Exe^{m'}_\alpha(\tau)$ is defined; on the other hand, it implies that $Lbl_\alpha(\langle m', h'\rangle) = Lbl_\alpha(\langle m, h\rangle) = Lbl(Exe^m_\alpha(\tau))$. Now, condition (LE) of $\mathcal{M}$ ensures that $Lbl(Exe^m_\alpha(\tau)) = \tau$. Therefore, $Lbl_\alpha(\langle m', h'\rangle) = \tau$, which by condition (EL) of $\mathcal{M}$ implies that $h' \in Exe^{m'}_\alpha(\tau)$. For the $\supseteq$ inclusion, assume that $h' \in Exe^{m'}_\alpha(\tau)$. This means that $Lbl_\alpha(\langle m', h'\rangle) = \tau$. Since $m \sim_\alpha m'$, constraint (UAAT) of $\mathcal{M}$ implies that $Exe^m_\alpha(\tau)$ is defined (and thus non-empty). For every $h \in Exe^m_\alpha(\tau)$, condition (LE) of $\mathcal{M}$ ensures that $Lbl_\alpha(\langle m, h\rangle) = \tau = Lbl_\alpha(\langle m', h'\rangle)$. Thus, for any $h \in Exe^m_\alpha(\tau)$, the fact that $m \sim_\alpha m'$ implies that $\langle m, h\rangle \sim'_\alpha \langle m', h'\rangle$. Therefore, $h' \in [Exe^m_\alpha(\tau)]^{m'}_\alpha$.

(iv) For the left-to-right direction, we work by contraposition. Assume that $Lbl(L) \notin TOptimal^m_\alpha$. Then there exists $\tau \in Tps^m_\alpha$ such that $Lbl(L) \prec_H \tau$. This means that ($\star$) for all n such that $m \sim_\alpha n$, $Exe^n_\alpha(Lbl(L)) \leq Exe^n_\alpha(\tau)$, and that ($\star\star$) there exists $m' \in M$ such that $m \sim_\alpha m'$ and such that $Exe^{m'}_\alpha(\tau) \nleq Exe^{m'}_\alpha(Lbl(L))$. We show that $Exe^m_\alpha(\tau)$ is such that $L \prec_s Exe^m_\alpha(\tau)$. For this, we need to show that (a) for all n such that $m \sim'_\alpha n$, $[L]^n_\alpha \leq [Exe^m_\alpha(\tau)]^n_\alpha$, and that (b) there exists $n' \in M$ such that $m \sim'_\alpha n'$ and such that $[Exe^m_\alpha(\tau)]^{n'}_\alpha \nleq [L]^{n'}_\alpha$. For (a), let n be such that $m \sim'_\alpha n$. Point ii implies that $[L]^n_\alpha = Exe^n_\alpha(Lbl(L))$, and point iii implies that $Exe^n_\alpha(\tau) = [Exe^m_\alpha(\tau)]^n_\alpha$. Since point i implies that $m \sim_\alpha n$, ($\star$) gives that $[L]^n_\alpha \leq [Exe^m_\alpha(\tau)]^n_\alpha$, as desired. For (b), ($\star\star$) and points i, ii, and iii give that m' is such that $m \sim'_\alpha m'$ and such that $[Exe^m_\alpha(\tau)]^{m'}_\alpha \nleq [L]^{m'}_\alpha$. Therefore, $L \prec_s Exe^m_\alpha(\tau)$, so that $L \notin \mathbf{SOptimal}^m_\alpha$.

For the right-to-left direction, we also work by contraposition. Assume that $L \notin \mathbf{SOptimal}^m_\alpha$. Then there exists $L' \in \mathbf{Choice}^m_\alpha$ such that $L \prec_s L'$. This means that ($\star$) for all n such that $m \sim'_\alpha n$, $[L]^n_\alpha \leq [L']^n_\alpha$, and that ($\star\star$) there exists $m' \in M$ such that $m \sim'_\alpha m'$ and such that $[L']^{m'}_\alpha \nleq [L]^{m'}_\alpha$. We show that $Lbl(L')$ is such that $Lbl(L) \prec_H Lbl(L')$. For this, we need to show that (a) for all n such that $m \sim_\alpha n$, $Exe^n_\alpha(Lbl(L)) \leq Exe^n_\alpha(Lbl(L'))$, and that (b) there exists $m' \in M$ such that $m \sim_\alpha m'$ and such that $Exe^{m'}_\alpha(Lbl(L')) \nleq Exe^{m'}_\alpha(Lbl(L))$. For (a), let n be such that $m \sim_\alpha n$. Point i implies that $m \sim'_\alpha n$. Point ii implies both that $Exe^n_\alpha(Lbl(L)) = [L]^n_\alpha$ and that $Exe^n_\alpha(Lbl(L')) = [L']^n_\alpha$. Therefore, ($\star$) gives that $Exe^n_\alpha(Lbl(L)) \leq Exe^n_\alpha(Lbl(L'))$. For (b), ($\star\star$) and points i and ii give that m' is such that $m \sim_\alpha m'$ and such that $Exe^{m'}_\alpha(Lbl(L')) \nleq Exe^{m'}_\alpha(Lbl(L))$. Therefore, $Lbl(L) \prec_H Lbl(L')$, so that $Lbl(L) \notin TOptimal^m_\alpha$.

□

Theorem 4.25 (Correspondence). *Let $\mathcal{M}$ be a finite-choice Horty-like labelled* eaubt-*model, and let $\mathcal{M}'$ be its transform finite-choice* eaubt-*model. Let us redefine $\leq_s$ in $\mathcal{M}'$ so that, for $\alpha \in Ags$, $m \in M$, and $L, L' \in$* **Choice**$^m_\alpha$, *$L \leq_s L'$ iff for all m' such that $m \sim_\alpha m'$, $[L]^{m'}_\alpha \leq [L']^{m'}_\alpha$. Then, for every formula φ of $\mathcal{L}_H$, $\mathcal{M}, \langle m, h\rangle \models \varphi$ iff $\mathcal{M}', \langle m, h\rangle \models Tr(\varphi)$.*

Proof. We proceed by induction on the complexity of φ. The base case follows from Definition 4.23. The cases of Boolean connectives are standard. The cases of modal operators $\Box$, $[\alpha\ \texttt{stit}]$, and $\odot[\alpha\ \texttt{stit}]$ follow from Definition 4.23. Let us deal with the cases of the remaining modal operators:

- ("K_α") ($\Rightarrow$) Assume that $\mathcal{M}, \langle m, h\rangle \models K_\alpha\varphi$. We show that $\mathcal{M}', \langle m, h\rangle \models \Box K_\alpha Tr(\varphi)$. Take $h' \in H_m$, and let $\langle m'', h''\rangle$ be an index such that $\langle m, h'\rangle \sim'_\alpha \langle m'', h''\rangle$. We want to show that $\mathcal{M}', \langle m'', h''\rangle \models Tr(\varphi)$. The fact that

$\langle m,h'\rangle \sim'_\alpha \langle m'',h''\rangle$ implies, by definition of $\sim'_\alpha$, that $\langle m,h'\rangle \sim_\alpha \langle m'',h''\rangle$. Observe that reflexivity of $\sim_\alpha$ and constraint ($\texttt{C4}$) imply that $\langle m,h\rangle \sim_\alpha \langle m,h'\rangle$. Therefore, transitivity of $\sim_\alpha$ yields that $\langle m,h\rangle \sim_\alpha \langle m'',h''\rangle$. Our assumption then implies that $\mathcal{M},\langle m'',h''\rangle \models \varphi$. The induction hypothesis then gives that $\mathcal{M}',\langle m'',h''\rangle \models Tr(\varphi)$. Therefore, $\mathcal{M}',\langle m,h\rangle \models \Box K_\alpha Tr(\varphi)$. ($\Leftarrow$) Assume that $\mathcal{M}',\langle m,h\rangle \models \Box K_\alpha Tr(\varphi)$. We show that $\mathcal{M},\langle m,h\rangle \models K_\alpha\varphi$. Let $\langle m'',h''\rangle$ be an index such that $\langle m,h\rangle \sim_\alpha \langle m'',h''\rangle$. We want to show that $\mathcal{M},\langle m'',h''\rangle \models \varphi$. Constraint ($\texttt{UAAT}$) ensures that there exists $h' \in H_m$ such that $Lbl_\alpha(\langle m'',h''\rangle) = Lbl_\alpha(\langle m,h'\rangle)$. Observe, then, that the fact that $\langle m,h\rangle \sim_\alpha \langle m'',h''\rangle$ implies, with constraint ($\texttt{C4}$), that $\langle m,h'\rangle \sim_\alpha \langle m'',h''\rangle$ (since $h' \in H_m$). By definition of $\sim'_\alpha$, this last fact implies that $\langle m,h'\rangle \sim'_\alpha \langle m'',h''\rangle$. Since our assumption implies that $\mathcal{M}',\langle m,h'\rangle \models K_\alpha Tr(\varphi)$, we then get that $\mathcal{M}',\langle m'',h''\rangle \models Tr(\varphi)$. The induction hypothesis then gives that $\mathcal{M},\langle m'',h''\rangle \models \varphi$, and thus that $\mathcal{M},\langle m,h\rangle \models K_\alpha\varphi$.

- ("$[\alpha\ \texttt{kstit}]$") ($\Rightarrow$) Assume that $\mathcal{M},\langle m,h\rangle \models [\alpha\ \texttt{kstit}]\varphi$. We show that $\mathcal{M}',\langle m,h\rangle \models K_\alpha Tr(\varphi)$. Let $\langle m',h'\rangle$ be an index such that $\langle m,h\rangle \sim'_\alpha \langle m',h'\rangle$. We want to show that $\mathcal{M}',\langle m',h'\rangle \models Tr(\varphi)$. The fact that $\langle m,h\rangle \sim'_\alpha \langle m',h'\rangle$ implies by definition of $\sim'_\alpha$ that $\langle m,h\rangle \sim_\alpha \langle m',h'\rangle$ and that $Lbl_\alpha(\langle m,h\rangle) = Lbl_\alpha(\langle m',h'\rangle)$. Thus, our assumption implies that $\mathcal{M},\langle m',h'\rangle \models \varphi$. The induction hypothesis then gives that $\mathcal{M}',\langle m',h'\rangle \models Tr(\varphi)$. Therefore, $\mathcal{M}',\langle m,h\rangle \models K_\alpha Tr(\varphi)$. ($\Leftarrow$) Assume that $\mathcal{M}',\langle m,h\rangle \models K_\alpha Tr(\varphi)$. We show that $\mathcal{M},\langle m,h\rangle \models [\alpha\ \texttt{kstit}]\varphi$. Let $\langle m',h'\rangle$ be an index such that $\langle m,h\rangle \sim_\alpha \langle m',h'\rangle$ and $Lbl_\alpha(\langle m,h\rangle) = Lbl_\alpha(\langle m',h'\rangle)$. The definition of $\sim'_\alpha$ ensures that $\langle m,h\rangle \sim'_\alpha \langle m',h'\rangle$. Our assumption then implies that $\mathcal{M}',\langle m,h'\rangle \models Tr(\varphi)$. The induction hypothesis then gives that $\mathcal{M},\langle m',h'\rangle \models \varphi$, and thus that $\mathcal{M},\langle m,h\rangle \models [\alpha\ \texttt{kstit}]\varphi$.

- ("$\odot[\alpha\ \texttt{kstit}]$") ($\Rightarrow$) Assume that $\mathcal{M},\langle m,h\rangle \models \odot[\alpha\ \texttt{kstit}]\varphi$. We show that $\mathcal{M}',\langle m,h\rangle \models \odot^{\mathcal{S}}_\alpha Tr(\varphi)$. Take $L \in \mathbf{SOptimal}^m_\alpha$, let m' be a moment such that $m \sim'_\alpha m'$, and take $h' \in [L]^{m'}_\alpha$. We want to show that $\mathcal{M}',\langle m',h'\rangle \models Tr(\varphi)$. By Lemma B.37 iv, the fact that $L \in \mathbf{SOptimal}^m_\alpha$ implies that $Lbl(L) \in TOptimal^m_\alpha$. The assumption that $\mathcal{M},\langle m,h\rangle \models \odot[\alpha\ \texttt{kstit}]\varphi$ and the fact that $m \sim_\alpha m'$ imply that, for all $h'' \in Exe^{m'}_\alpha(Lbl(L))$, $\mathcal{M},\langle m',h''\rangle \models \varphi$. But Lemma B.37 ii gives that $[L]^{m'}_\alpha = Exe^{m'}_\alpha(Lbl(L))$. Thus, the fact that we took $h' \in [L]^{m'}_\alpha$ implies that $\mathcal{M},\langle m',h'\rangle \models \varphi$. By induction hypothesis, then, $\mathcal{M}',\langle m',h'\rangle \models Tr(\varphi)$, so that $\mathcal{M}',\langle m,h\rangle \models \odot^{\mathcal{S}}_\alpha Tr(\varphi)$. ($\Leftarrow$) Assume that $\mathcal{M}',\langle m,h\rangle \models \odot^{\mathcal{S}}_\alpha Tr(\varphi)$. We show that $\mathcal{M},\langle m,h\rangle \models \odot[\alpha\ \texttt{kstit}]\varphi$. Take $\tau \in TOptimal^m_\alpha$, let m' be a moment such that $m \sim_\alpha m'$, and take $h' \in Exe^{m'}_\alpha(\tau)$. We want to show that $\mathcal{M},\langle m',h'\rangle \models \varphi$. By condition ($\texttt{LE}$) of $\mathcal{M}$, $Lbl(Exe^m_\alpha(\tau)) = \tau \in TOptimal^m_\alpha$. By Lemma B.37 iv, this implies that $Exe^{m}_{\alpha}(\tau) \in \mathbf{SOptimal}^{m}_{\alpha}$. The assumption that $\mathcal{M}', \langle m,h\rangle \models \odot^{S}_{\alpha} Tr(\varphi)$ and the fact that $m \sim'_{\alpha} m'$ imply that, for every $h'' \in [Exe^{m}_{\alpha}(\tau)]^{m'}_{\alpha}$, $\mathcal{M}', \langle m',h''\rangle \models Tr(\varphi)$. But Lemma B.37 iii gives that $Exe^{m'}_{\alpha}(\tau) = [Exe^{m}_{\alpha}(\tau)]^{m'}_{\alpha}$. Thus, the fact that we took $h' \in Exe^{m'}_{\alpha}(\tau)$ implies that $\mathcal{M}', \langle m',h'\rangle \models Tr(\varphi)$. By induction hypothesis, then, $\mathcal{M}, \langle m',h'\rangle \models \varphi$, so that $\mathcal{M}, \langle m,h\rangle \models \odot[\alpha\,\texttt{kstit}]\varphi$.

□

Appendix C Metalogic Results for *EAUST*

C.1 Equivalence of Proof Systems for Objective Ought-to-Do's

Proposition C.38 (Equivalence of Λ_O to Murakami's (2004) proof system)**.** *Definition 4.27's schemata* (SET)*,* ($A1$) *–* ($A4$)*,* (Oic)*, and* (IA) *are jointly equivalent to the schemata that Murakami used to axiomatize Horty's (2001) logic of act-utilitarian ought-to-do.*

Proof. Instead of ($A3$), Murakami used schema (a) $\Box \odot_{\alpha} \varphi \vee \Box\neg \odot_{\alpha} \varphi$, and instead of ($A4$), Murakami used schema (b) $\Box([\alpha]\varphi \to [\alpha]\psi) \to (\odot_{\alpha}\varphi \to \odot_{\alpha}\psi)$. With the exception of the schema that Murakami used to characterize syntactically that the cardinality of available choices is finite (see the discussion of schemata ($AgsPC_n$) and (APC_n) in Chapter 3's Section 3.4, as well as Footnote 34, p. 105), each of the remaining schemata in her system is logically equivalent to one in Λ_O.

Now, Murakami's schema (a) is derivable in Λ_O according to the following arguments: first of all, observe that formula ($\star$) $\neg \odot_{\alpha} \varphi \to \Box\neg \odot_{\alpha} \varphi$ is derivable in Λ_O. A derivation is as follows, where 'c.p.' abbreviates 'contrapositive' and 'Subs.' abbreviates 'Substitution':

1. $\vdash_{\Lambda_O}$	$\Diamond\neg \odot_{\alpha} \varphi \to \Box\Diamond\neg \odot_{\alpha} \varphi$	Subs. of (5) for □
2. $\vdash_{\Lambda_O}$	$\Diamond\Box \odot_{\alpha} \varphi \to \Box \odot_{\alpha} \varphi$	C.p. of 1
3. $\vdash_{\Lambda_O}$	$\odot_{\alpha}\varphi \to \Box \odot_{\alpha} \varphi$	($A3$)
4. $\vdash_{\Lambda_O}$	$\Diamond \odot_{\alpha} \varphi \to \Diamond\Box \odot_{\alpha} \varphi$	3, modal logic
5. $\vdash_{\Lambda_O}$	$\Diamond \odot_{\alpha} \varphi \to \Box \odot_{\alpha} \varphi$	4, 2, prop. logic
6. $\vdash_{\Lambda_O}$	$\Diamond \odot_{\alpha} \varphi \to \odot_{\alpha}\varphi$	(T) for □, 5, prop. logic
7. $\vdash_{\Lambda_O}$	$\neg \odot_{\alpha} \varphi \wedge \Diamond \odot_{\alpha} \varphi \to \neg \odot_{\alpha} \varphi \wedge \odot_{\alpha}\varphi$	6, prop. logic
8. $\vdash_{\Lambda_O}$	$\neg \odot_{\alpha} \varphi \wedge \Diamond \odot_{\alpha} \varphi \to \bot$	7, prop. logic
9. $\vdash_{\Lambda_O}$	$\neg \odot_{\alpha} \varphi \to \Box\neg \odot_{\alpha} \varphi$	8, prop. & modal logic.

A derivation for schema (a) is as follows, then:

1. $\vdash_{\Lambda_O}$	$\odot_\alpha\varphi \vee \neg \odot_\alpha \varphi$	Prop. logic
2. $\vdash_{\Lambda_O}$	$\odot_\alpha\varphi \vee \neg \odot_\alpha \varphi \rightarrow \Box \odot_\alpha \varphi \vee \Box\neg \odot_\alpha \varphi$	$(A3), (\star)$, prop. logic
3. $\vdash_{\Lambda_O}$	$\Box \odot_\alpha \varphi \vee \Box\neg \odot_\alpha \varphi$	2, 1, *Modus Ponens*.

Similarly, Murakami's schema (b) is derivable in Λ_O, according to the following derivation:

1. $\vdash_{\Lambda_O}$	$\Box([\alpha]\varphi \rightarrow [\alpha]\psi) \rightarrow \odot_\alpha([\alpha]\varphi \rightarrow [\alpha]\psi)$	Subs. of $(A2)$
2. $\vdash_{\Lambda_O}$	$\odot_\alpha([\alpha]\varphi \rightarrow [\alpha]\psi) \rightarrow (\odot_\alpha[\alpha]\varphi \rightarrow \odot_\alpha[\alpha]\psi)$	Subs. of (K) for $\odot_\alpha$
3. $\vdash_{\Lambda_O}$	$((\odot_\alpha[\alpha]\varphi \rightarrow \odot_\alpha[\alpha]\psi) \wedge \odot_\alpha[\alpha]\varphi) \rightarrow \odot_\alpha[\alpha]\psi$	Prop. logic
4. $\vdash_{\Lambda_O}$	$\odot_\alpha[\alpha]\psi \rightarrow \odot_\alpha\psi$	Subs. of (T) for $[\alpha]$, modal logic
5. $\vdash_{\Lambda_O}$	$((\odot_\alpha[\alpha]\varphi \rightarrow \odot_\alpha[\alpha]\psi) \wedge \odot_\alpha[\alpha]\varphi) \rightarrow \odot_\alpha\psi$	3, 4, prop. logic
6. $\vdash_{\Lambda_O}$	$\odot_\alpha\varphi \rightarrow \odot_\alpha([\alpha]\varphi)$	$(A6)$
7. $\vdash_{\Lambda_O}$	$((\odot_\alpha[\alpha]\varphi \rightarrow \odot_\alpha[\alpha]\psi) \wedge \odot_\alpha\varphi) \rightarrow \odot_\alpha\psi$	5, 6, prop. logic
8. $\vdash_{\Lambda_O}$	$(\Box([\alpha]\varphi \rightarrow [\alpha]\psi) \wedge \odot_\alpha\varphi) \rightarrow \odot_\alpha\psi$	1, 2, 7, prop. logic.

The other way around, schema $(A3)$ is derivable in Murakami's proof system, according to the following derivation:

1. $\vdash_{\Lambda_O}$	$\odot_\alpha\varphi \rightarrow (\Box \odot_\alpha \varphi \vee \Box\neg \odot_\alpha \varphi)$	Prop. logic
2. $\vdash_{\Lambda_O}$	$(\odot_\alpha\varphi \rightarrow (\Box \odot_\alpha \varphi \vee \Box\neg \odot_\alpha \varphi)) \rightarrow$	
	$(\odot_\alpha\varphi \rightarrow \Box \odot_\alpha \varphi) \vee (\odot_\alpha\varphi \rightarrow \Box\neg \odot_\alpha \varphi)$	Prop. logic
3. $\vdash_{\Lambda_O}$	$(\odot_\alpha\varphi \rightarrow \Box \odot_\alpha \varphi) \vee (\odot_\alpha\varphi \rightarrow \Box\neg \odot_\alpha \varphi)$	1, 2, *Modus Ponens*
4. $\vdash_{\Lambda_O}$	$(\odot_\alpha\varphi \rightarrow \Box\neg \odot_\alpha \varphi) \rightarrow \bot$	Subs. of (T) for $\Box$, Prop. logic
5. $\vdash_{\Lambda_O}$	$\odot_\alpha\varphi \rightarrow \Box \odot_\alpha \varphi$	3, 4, *Modus Ponens*.

Similarly, schema $(A6)$ is derivable in Murakami's proof system, according to the following derivation, where 'Nec.' abbreviates 'Necessitation':

1. $\vdash_{\Lambda_O}$	$[\alpha]\varphi \rightarrow [\alpha][\alpha]\varphi$	Subs. of (4) for $[\alpha]$
2. $\vdash_{\Lambda_O}$	$\Box([\alpha]\varphi \rightarrow [\alpha][\alpha]\varphi)$	1, Nec. for $\Box$
3. $\vdash_{\Lambda_O}$	$\odot_\alpha\varphi \rightarrow \odot_\alpha([\alpha]\varphi)$	Schema (b), 2, *Modus Ponens*.

Thus, Λ_O is equivalent to Murakami's (2004) proof system that does not include the schema that Murakami used to characterize syntactically that the cardinality of available choices is finite. □

C.2 Soundness for Subjective Ought-to-Do's

Observation C.39. *Objective and subjective ought-to-do's are consistent. In other words, the formulas (a)* $\odot_\alpha\varphi \to \neg \odot^{\mathcal{S}}_\alpha \neg\varphi$ *and (b)* $\odot^{\mathcal{S}}_\alpha\varphi \to \neg \odot_\alpha \neg\varphi$ *are valid with respect to the class of* eaubt-*models.*

Proof. By Substitution and contraposition, the validity of one of these formulas implies the validity of the other. Therefore, let us show that (b) is valid.

Let $\mathcal{M}$ be an *eaubt*-model. First of all, let us show that, for all $L, L' \in \mathbf{Choice}^m_\alpha$, if $L \preceq_s L'$, then $L \preceq L'$. Take $L, L' \in \mathbf{Choice}^m_\alpha$. If $L \preceq_s L'$, then, for each m' such that $m \sim_\alpha m'$, $\mathbf{Value}(h) \leq \mathbf{Value}(h')$ for every $h \in [L]^{m'}_\alpha, h' \in [L']^{m'}_\alpha$. Reflexivity of $\sim_\alpha$ implies both that $m \sim_\alpha m'$ and that $L \subseteq [L]^m_\alpha$ and $L' \subseteq [L']^m_\alpha$. Therefore, for all $h'' \in L$ and $h''' \in L'$, $\mathbf{Value}(h'') \leq \mathbf{Value}(h''')$, which implies that $L \preceq L'$.

Now, let $\langle m, h\rangle$ be an index. Assume for a contradiction that ($\star$) $\mathcal{M}, \langle m, h\rangle \models \odot^{\mathcal{S}}_\alpha\varphi$ and that ($\star\star$) $\mathcal{M}, \langle m, h\rangle \models \odot_\alpha\neg\varphi$. On the one hand, assumption ($\star$) implies that there is $L_* \in \mathbf{Choice}^m_\alpha$ such that $L_* \subseteq \left|\varphi\right|^m$. Thus, assumption ($\star\star$) yields that there is $L'_* \in \mathbf{Choice}^m_\alpha$ such that $L_* \prec L'_*$ and, if $N = L'_*$ or $L'_* \preceq N$, then $N \subseteq |\neg\varphi|^m$. In particular, $L'_* \subseteq |\neg\varphi|^m$. Assumption ($\star$) then implies that there is $L''_* \in \mathbf{Choice}^m_\alpha$ such that $L'_* \prec_s L''_*$ and, if $N = L''_*$ or $L''_* \preceq_s N$, then $N \subseteq [N]^m_\alpha \subseteq \left|\varphi\right|^m$. In particular, $L''_* \subseteq \left|\varphi\right|^m$. On the other hand, by the first observation in the proof, the fact that $L'_* \prec_s L''_*$ implies that $L'_* \preceq L''_*$, so that assumption ($\star\star$) yields that $L''_* \subseteq |\neg\varphi|^m$, which contradicts the previously shown fact that $L''_* \subseteq \left|\varphi\right|^m$. Thus, $\mathcal{M}, \langle m, h\rangle \models \odot^{\mathcal{S}}_\alpha\varphi \to \neg \odot_\alpha \neg\varphi$ for every index $\langle m, h\rangle$, so that $\odot^{\mathcal{S}}_\alpha\varphi \to \neg \odot_\alpha \neg\varphi$ is indeed valid. □

Proposition C.40 (Soundness of Λ_S)**.** *The system* Λ_S *is sound with respect to the class of* eaubt-*models.*

Proof. Let $\left\langle M, \sqsubset, Ags, \mathbf{Choice}, \{\sim_\alpha\}_{\alpha\in Ags}, \mathbf{Value}\right\rangle$ be an *eaubt*-frame. Let $\mathcal{V}$ be any valuation function, and let $\mathcal{M}$ be the *eaubt*-model that results from adding $\mathcal{V}$ to the *eaubt*-frame above. The **S5** schemata for $\Box$, $[\alpha]$, and K_α, as well as schemata (SET), (IA), and ($A5$), are shown to be valid straightforwardly. Since they involve some novelty, I include the detailed proofs for the validity of schemata (OAC), ($Unif - H$), ($A6$), (SuN), ($s.Oic$), and ($s.Cl$) below.

- To see that $\mathcal{M} \models (OAC)$, take $\langle m, h\rangle$ such that $\mathcal{M}, \langle m, h\rangle \models K_\alpha\varphi$. Take $h' \in \mathbf{Choice}^m_\alpha(h)$. Frame condition (OAC) implies that $\langle m, h\rangle \sim_\alpha \langle m, h'\rangle$. The assumption that $\mathcal{M}, \langle m, h\rangle \models K_\alpha\varphi$ then implies that $\mathcal{M}, \langle m, h'\rangle \models \varphi$. Therefore, for any $h' \in \mathbf{Choice}^m_\alpha(h)$, $\mathcal{M}, \langle m, h'\rangle \models \varphi$, which implies that $\mathcal{M}, \langle m, h\rangle \models [\alpha]\varphi$.
- To see that $\mathcal{M} \models (Unif - H)$, take $\langle m, h\rangle$ such that $\mathcal{M}, \langle m, h\rangle \models \Diamond K_\alpha\varphi$. Let $\langle m', h'\rangle$ be an index such that $\langle m, h\rangle \sim_\alpha \langle m', h'\rangle$. We want to show that

$\mathcal{M},\langle m',h'\rangle \models \Diamond\varphi$. The fact that $\mathcal{M},\langle m,h\rangle \models \Diamond K_\alpha\varphi$ implies that there exists $h_* \in H_m$ such that ($\star$) $\mathcal{M},\langle m,h_*\rangle \models K_\alpha\varphi$. Frame condition ($\mathtt{Unif-H}$) implies that there exists $h'_* \in H_{m'}$ such that $\langle m,h_*\rangle \sim_\alpha \langle m',h'_*\rangle$. With ($\star$), this last fact implies that $\mathcal{M},\langle m',h'_*\rangle \models \varphi$, which in turn implies that $\mathcal{M},\langle m',h'\rangle \models \Diamond\varphi$. Therefore, $\mathcal{M},\langle m,h\rangle \models K_\alpha\Diamond\varphi$.

- To see that $\mathcal{M} \models (A6)$, take $\langle m,h\rangle$ such that $\mathcal{M},\langle m,h\rangle \models \odot^{S}_{\alpha}\varphi$. We want to show that, for every $L \in \mathbf{Choice}^m_\alpha$ such that $[L]^{m'} \nsubseteq \left|K_\alpha\varphi\right|^{m'}$ (for some m' such that $m \sim_\alpha m'$), there is $L' \in \mathbf{Choice}^m_\alpha$ such that $L \prec_s L'$ and, if $L'' = L'$ or $L' \preceq_s L''$, then $[L'']^{m''}_\alpha \subseteq \left|K_\alpha\varphi\right|^{m''}$ for every m'' such that $m \sim_\alpha m''$. Take $L \in \mathbf{Choice}^m_\alpha$ such that there exists $m' \in M$ such that $m \sim_\alpha m'$ and $[L]^{m'} \nsubseteq |K_\alpha\varphi|^{m'}$. This implies that $[L]^{m'''} \nsubseteq \left|\varphi\right|^{m'''}$ for some m''' such that $m' \sim_\alpha m'''$. Now, transitivity of $\sim_\alpha$ implies that $m \sim_\alpha m'''$. Therefore, the assumption that $\mathcal{M},\langle m,h\rangle \models \odot^{S}_{\alpha}\varphi$ implies that there is $L' \in \mathbf{Choice}^m_\alpha$ such that $L \prec_s L'$ and, if $L'' = L'$ or $L' \preceq_s L''$, then $[L'']^{m''}_\alpha \subseteq \left|\varphi\right|^{m''}$ for every m'' such that $m \sim_\alpha m''$. By definition of epistemic clusters (Definition 4.19) and transitivity of $\sim_\alpha$, this last clause implies that if $L'' = L'$ or $L' \preceq_s L''$ then $[L'']^{m''}_\alpha \subseteq \left|K_\alpha\varphi\right|^{m''}$ for every m'' such that $m \sim_\alpha m''$. Thus, L' attests to the fact that $\mathcal{M},\langle m,h\rangle \models \odot^{S}_{\alpha}(K_\alpha\varphi)$.

- To see that $\mathcal{M} \models (SuN)$, take $\langle m,h\rangle$ such that $\mathcal{M},\langle m,h\rangle \models K_\alpha\Box\varphi$. Take $L \in \mathbf{Choice}^m_\alpha$, and let $m' \in M$ be such that $m \sim_\alpha m'$ (which means that there exist $j \in H_m$, $j' \in H_{m'}$ such that $\langle m,j\rangle \sim_\alpha \langle m',j'\rangle$). Condition ($\mathtt{Unif-H}$) ensures that there exists $h' \in H_{m'}$ such that $\langle m,h\rangle \sim_\alpha \langle m',h'\rangle$. The assumption that $\mathcal{M},\langle m,h\rangle \models K_\alpha\Box\varphi$ then implies that $\mathcal{M},\langle m',h'\rangle \models \Box\varphi$. Thus, for any $h'' \in [L]^{m'}_\alpha$, the fact that $h'' \in H_{m'}$ yields that $\mathcal{M},\langle m',h''\rangle \models \varphi$. Therefore, for all $L \in \mathbf{Choice}^m_\alpha$ and m' such that $m \sim_\alpha m'$, $[L]^{m'}_\alpha \subseteq \left|\varphi\right|^{m'}$, which vacuously implies that $\mathcal{M},\langle m,h\rangle \models \odot^{S}_{\alpha}\varphi$.

- To see that $\mathcal{M} \models (s.Oic)$, take $\langle m,h\rangle$ such that $\mathcal{M},\langle m,h\rangle \models \odot^{S}_{\alpha}\varphi$. This implies that there exists $L \subseteq H_m$ such that $[L]^{m''}_\alpha \subseteq \left|\varphi\right|^{m''}$ for every $m'' \in M$ such that $m \sim_\alpha m''$. Since $\sim_\alpha$ is reflexive, $[L]^{m}_\alpha \subseteq \left|\varphi\right|^{m}$. Now, take $h_0 \in L$. Let $\langle m',h'\rangle$ be an index such that $\langle m,h_0\rangle \sim_\alpha \langle m',h'\rangle$. From the definition of epistemic clusters (Definition 4.19), $h' \in [L]^{m'}_\alpha$, so the fact that $[L]^{m'}_\alpha \subseteq \left|\varphi\right|^{m'}$ implies that $\mathcal{M},\langle m',h'\rangle \models \varphi$. Therefore, history $h_0 \in H_m$ is such that, for every $\langle m',h'\rangle$ with $\langle m,h_0\rangle \sim_\alpha \langle m',h'\rangle$, $\mathcal{M},\langle m',h'\rangle \models \varphi$. This means that $\mathcal{M},\langle m,h_0\rangle \models K_\alpha\varphi$, which implies that $\mathcal{M},\langle m,h\rangle \models \Diamond K_\alpha\varphi$.

- To see that $\mathcal{M} \models (s.Cl)$, take $\langle m_*, h_* \rangle$ such that $\mathcal{M}, \langle m_*, h_* \rangle \models \odot^S_\alpha \varphi$. Let $\langle m, j \rangle$ be such that $\langle m_*, h_* \rangle \sim_\alpha \langle m, j \rangle$. Take $h \in H_m$. We want to show that, for every $L \in \mathbf{Choice}^m_\alpha$ such that $[L]^{m'} \nsubseteq \left|\varphi\right|^{m'}$ (for some m' such that $m \sim_\alpha m'$), there is $L' \in \mathbf{Choice}^m_\alpha$ such that $L \prec_s L'$ and, if $L'' = L'$ or $L' \preceq_s L''$, then $[L'']^{m''}_\alpha \subseteq \left|\varphi\right|^{m''}$ for every m'' such that $m \sim_\alpha m''$. Take $L \in \mathbf{Choice}^m_\alpha$ such that there exists $m' \in M$ such that $m \sim_\alpha m'$ and $[L]^{m'} \nsubseteq \left|\varphi\right|^{m'}$. Let N_L be an action in $\mathbf{Choice}^{m_*}_\alpha$ such that $N_L \subseteq [L]^{m_*}_\alpha$, where we know that such an action exists in virtue of $(\mathtt{Unif-H})$ and $(\mathtt{OAC})$. Notice that transitivity of $\sim_\alpha$ entails that $[N_L]^o_\alpha = [L]^o_\alpha$ for any moment o, so that $[N_L]^{m'}_\alpha \nsubseteq \left|\varphi\right|^{m'}$. Since $\mathcal{M}, \langle m_*, h_* \rangle \models \odot^S_\alpha \varphi$, there must exist $N \in \mathbf{Choice}^{m_*}_\alpha$ such that $N_L \prec_s N$ and, if $N' = N$ or $N \preceq_s N'$, then $[N']^{m''}_\alpha \subseteq \left|\varphi\right|^{m''}$ for every m'' such that $m_* \sim_\alpha m''$. Now, let L_N be an action in $\mathbf{Choice}^m_\alpha$ such that $L_N \subseteq [N]^m_\alpha$ (which implies that $[L_N]^o_\alpha = [N]^o_\alpha$ for any moment o). We claim that $L \prec_s L_N$, and show our claim with the following argument: let $m'' \in M$ be such that $m \sim_\alpha m''$, and take $S \in \mathbf{State}^{m''}_\alpha$; on the one hand, $(\star)$ $[L]^{m''}_\alpha \cap S = [N_L]^{m''}_\alpha \cap S \leq [N]^{m''}_\alpha \cap S = [L_N]^{m''}_\alpha \cap S$; on the other hand, we know that there exist a moment m''' and a state $S_0 \in \mathbf{State}^{m'''}_\alpha$ such that $m_* \sim_\alpha m'''$ and such that $[N]^{m'''}_\alpha \cap S_0 \nleq [N_L]^{m'''}_\alpha \cap S_0$; therefore, $(\star\star)$ $[L_N]^{m'''}_\alpha \cap S_0 = [N]^{m'''}_\alpha \cap S_0 \nleq [N_L]^{m'''}_\alpha \cap S_0 = [L]^{m'''}_\alpha \cap S_0$. Together, $(\star)$ and $(\star\star)$ entail that $L \prec_s L_N$, proving our claim. Now, let $L'' \in \mathbf{Choice}^m_\alpha$ be such that $L'' = L_N$ or $L_N \preceq_s L''$. If $L'' = L_N$, then $[L'']^{m''}_\alpha = [N]^{m''}_\alpha \subseteq \left|\varphi\right|^{m''}$ for every m'' such that $m \sim_\alpha m''$. If $L_N \prec_s L''$, then an argument similar to the one used to show that our claim was true renders that there is an action $N_{L''} \in \mathbf{Choice}^{m_*}_\alpha$ such that $N_{L''} \subseteq [L'']^{m_*}_\alpha$ and $N \preceq_s N_{L''}$. Thus, $[L'']^{m''}_\alpha = [N_{L''}]^{m''}_\alpha \subseteq \left|\varphi\right|^{m''}$. With this, we have shown that $\mathcal{M}, \langle m, h \rangle \models \odot^S_\alpha \varphi$ for every $h \in H_m$, so that $\mathcal{M}, \langle m, j \rangle \models \Box \odot^S_\alpha \varphi$. But $\langle m, j \rangle$ was an arbitrary index such that $\langle m_*, h_* \rangle \sim_\alpha \langle m, j \rangle$. Thus, $\mathcal{M}, \langle m_*, h_* \rangle \models K_\alpha \Box \odot^S_\alpha \varphi$.

- It is clear that the rules of inference *Modus Ponens*, Substitution, and Necessitation for the modal operators all preserve validity.

Therefore, Λ_S is sound with respect to the class of *eaubt*-models. □

C.3 Completeness for Subjective Ought-to-Do's

C.3.1 From Kripke Models to Stit Models

To prove completeness of Λ_S with respect to *eaubt*-models, we will first prove completeness with respect to a class of Kripke models. The reason is that there exists a truth-preserving correspondence between this class and a sub-class of *eaubt*-models. Below, I define said class of Kripke models and prove such a truth-preserving correspondence.

Definition C.41 (Kripke-*eaus*-frames & models). *A tuple*

$$\left\langle W, Ags, R_{\Box}, \mathtt{Choice}, \{\approx_\alpha\}_{\alpha \in Ags}, \mathtt{Value} \right\rangle$$

*is called a Kripke-*eaus*-frame iff*

- $\langle W, Ags, R_{\Box}, \mathtt{Choice}\rangle$ *is a Kripke-*stit*-frame (Definition 2.9).*

  Recall that, for $w \in W$ *and* $v \in \overline{w}$*, the class of* v *in the partition* $\mathtt{Choice}^{\overline{w}}_\alpha$ *is denoted by* $\mathtt{Choice}^{\overline{w}}_\alpha(v)$*. Now, for* $\beta \in Ags$ *and* $w \in W$*,* $\mathtt{State}^{\overline{w}}_\beta := \left\{S \subseteq \overline{w}; S = \bigcap_{\alpha \in Ags - \{\beta\}} s(\alpha), \text{ for } s \in \mathtt{Select}^{\overline{w}}\right\}$*, where* $\mathtt{Select}^{\overline{w}}$ *denotes the set of all selection functions at* $\overline{w}$ *(i.e., functions that assign to each* α *a member of* $\mathtt{Choice}^{\overline{w}}_\alpha$*).*

- *For all* $\alpha \in Ags$*,* $\approx_\alpha$ *is an (epistemic) equivalence relation on* W*. The following conditions must be satisfied:*

  - $(\mathtt{OAC})_{\mathtt{K}}$ *For all* $\alpha \in Ags$*,* $w \in W$*, and* $v \in \overline{w}$*,* $v \approx_\alpha u$ *for every* $u \in \mathtt{Choice}^{\overline{w}}_\alpha(v)$*.*
  - $(\mathtt{Unif-H})_{\mathtt{K}}$ *For all* $\alpha \in Ags$*, if* $v, u \in W$ *are such that* $v \approx_\alpha u$*, then for all* $v' \in \overline{v}$ *there exists* $u' \in \overline{u}$ *such that* $v' \approx_\alpha u'$*.*

  For $w, v \in W$*, I write* $\overline{w} \approx_\alpha \overline{v}$ *iff there exist* $w' \in \overline{w}$ *and* $v' \in \overline{v}$ *such that* $w' \approx_\alpha v'$*. For* $w, v \in W$ *such that* $\overline{w} \approx_\alpha \overline{v}$ *and* $L \in \mathtt{Choice}^{\overline{w}}_\alpha$*, L's epistemic cluster at* $\overline{v}$ *is the set* $[\![L]\!]^{\overline{v}}_\alpha := \{u \in \overline{v};\ \text{there is } o \in L \ \text{ such that } \ o \approx_\alpha u\}$*.*

- $\mathtt{Value}$ *is a function that assigns to each* $w \in W$ *a value in* $\mathbb{R}$*, representing the (deontic) utility of* w*. This function is used to define a subjective ordering* $\preceq_s$ *just as done in* eaubt*-models. For all* $\alpha \in Ags$ *and* $w \in W$*, this ordering leads to the corresponding notion of* $\mathtt{SOptimal}^{\overline{w}}_\alpha$*. Formally, for* $\alpha \in Ags$ *and* $w \in W$*, one first defines a general ordering* $\leq$ *on* 2^W *by the rule:* $X \leq Y$ *iff* $\mathtt{Value}(w) \leq \mathtt{Value}(w')$ *for all* $w \in X$ *and* $w' \in Y$*. The subjective dominance ordering* $\preceq_s$ *is then defined on* $\mathtt{Choice}^{\overline{w}}_\alpha$ *by the rule:* $L \preceq_s L'$ *iff* $[\![L]\!]^{\overline{v}}_\alpha \cap S \leq [\![L']\!]^{\overline{v}}_\alpha \cap S$ *for every* v *such that* $w \approx_\alpha v$ *and every* $S \in \mathtt{State}^{\overline{v}}_\alpha$*. I write* $L \prec_s L'$ *iff* $L \preceq_s L'$ *and* $L' \npreceq_s L$*, so that* $\mathtt{SOptimal}^{\overline{w}}_\alpha := \left\{L \in \mathtt{Choice}^{\overline{w}}_\alpha;\ \text{there is no } L' \in \mathtt{Choice}^{\overline{w}}_\alpha \text{ s. t. } L \prec_s L'\right\}$*.*

*A Kripke-*eaus*-model* $\mathcal{M}$ *consists of the tuple that results from adding a valuation function* $\mathcal{V}$ *to a Kripke-*eaus*-frame, where* $\mathcal{V} : P \to 2^W$ *assigns to each atomic proposition a set of worlds (recall that* P *is the set of propositions in* $\mathcal{L}_S$*).*

Kripke-*eaus*-models allow us to evaluate the formulas of $\mathcal{L}_\mathsf{S}$ with semantics that are analogous to the ones provided for *eaubt*-models (see Definition 4.21):

Definition C.42 (Evaluation rules on Kripke models)**.** *Let $\mathcal{M}$ be a Kripke-*eaus*-model, the semantics on $\mathcal{M}$ for the formulas of $\mathcal{L}_\mathsf{S}$ are defined recursively by the following truth conditions, evaluated at world w:*

$$\begin{array}{lll}
\mathcal{M}, w \models p & \textit{iff} & w \in \mathcal{V}(p) \\
\mathcal{M}, w \models \neg\varphi & \textit{iff} & \mathcal{M}, w \not\models \varphi \\
\mathcal{M}, w \models \varphi \wedge \psi & \textit{iff} & \mathcal{M}, w \models \varphi \textit{ and } \mathcal{M}, w \models \psi \\
\mathcal{M}, w \models \Box\varphi & \textit{iff} & \textit{for all } v \in \overline{w}, \mathcal{M}, v \models \varphi \\
\mathcal{M}, w \models [\alpha]\varphi & \textit{iff} & \textit{for all } v \in \mathtt{Choice}^{\overline{w}}_{\alpha}(w), \mathcal{M}, v \models \varphi \\
\mathcal{M}, w \models K_\alpha\varphi & \textit{iff} & \textit{for all } v \textit{ s. t. } w \approx_\alpha v, \mathcal{M}, v \models \varphi \\
\mathcal{M}, w \models \odot^{\mathcal{S}}_{\alpha}\varphi & \textit{iff} & \textit{for all } L \in \mathtt{Choice}^{\overline{w}}_{\alpha} \textit{ s. t. } \mathcal{M}, v \not\models \varphi \textit{ for some } w' \textit{ s. t. } w \approx_\alpha w' \\
 & & \textit{and some } v \in [\![L]\!]^{w'}_{\alpha}\,, \textit{ there is } L' \in \mathtt{Choice}^{\overline{w}}_{\alpha} \textit{ s. t. } L \lhd_s L' \\
 & & \textit{and, if } L'' = L' \textit{ or } L' \leq_s L'', \textit{then } \mathcal{M}, w''' \models \varphi \textit{ for every } w'' \\
 & & \textit{s. t. } \overline{w} \approx_\alpha \overline{w''} \textit{ and every } w''' \in [\![L'']\!]^{w''}_{\alpha}\,.
\end{array}$$

Satisfiability, validity, and general validity are defined as usual. I write $\left|\varphi\right|$ to refer to the set $\{w \in W; \mathcal{M}, w \models \varphi\}$.

Importantly, Kripke-*eaus*-models can be used for constructing *eaubt*-models such that both satisfy the same formulas of $\mathcal{L}_\mathsf{S}$, according to the following definition and propositions.

Definition C.43 (Associated *eaubt*-frame)**.** *Let*

$$\mathcal{F} = \left\langle W, Ags, R_\Box, \mathtt{Choice}, \{\approx_\alpha\}_{\alpha \in Ags}, \mathtt{Value} \right\rangle$$

*be a Kripke-*eaus*-frame. The tuple $\mathcal{F}^T := \left\langle M_W, \sqsubset, Ags, \mathbf{Choice}, \{\sim_\alpha\}_{\alpha \in Ags}, \mathbf{Value} \right\rangle$ is called the* eaubt*-frame associated with $\mathcal{F}$ iff*

- *$M_W := W \cup \{\overline{w}; w \in W\} \cup \{W\}$, and $\sqsubset$ is a relation on M_W, defined as the transitive closure of the union $\{(\overline{w}, v); w \in W \text{ and } v \in \overline{w}\} \cup \{(W, \overline{w}); w \in W\}$.*

  Observe that $\sqsubset$ is a strict partial order on M_W that straightforwardly satisfies no backward branching. Since the tuple $\langle M_W, \sqsubset\rangle$ is thus a tree, let us refer to the maximal $\sqsubset$-chains in M_W as histories, and let us denote by H_W the set of all histories of M_W. Observe that the definition of $\sqsubset$ yields that there is a bijective correspondence between W and H_W. For $v \in W$, let h_v be the history $\{W, \overline{v}, v\}$. Thus, for all $o \in W$, $o \in h_v$ iff $o = v$. Therefore, each history in H_W can be identified *using the world at its terminal node. Consequently, for all $w \in W$, if $H_{\overline{w}}$ denotes the set of histories passing through $\overline{w}$, then $H_{\overline{w}} = \{h_v; v \in \overline{w}\}$—since $\overline{w} \in h_v$ iff $v \in \overline{w}$. Observe, then, that $H_W = \{h_v; v \in W\}$.*

- *For $B \in 2^W$, let B^T denote the set $\{h_v;\ v \in B\}$. We then define* **Choice** *as a function on $Ags \times M_W$ given by the rules: for $\alpha \in Ags$,* **Choice**$(\alpha, W) = \{H_W\}$*; for $\alpha \in Ags$ and $w \in W$,* **Choice**$(\alpha, \overline{w}) = \left\{C_\alpha^T; C_\alpha \in \texttt{Choice}_\alpha^{\overline{w}}\right\}$*; for $\alpha \in Ags$ and $v \in W$,* **Choice**$(\alpha, v) = \{\{h_v\}\}$*. To keep notation consistent, the sets of the form* **Choice**$(\alpha, \overline{w})$ *are denoted by* **Choice**$_\alpha^{\overline{w}}$*, and the choice-cell of a given h_v in* **Choice**$_\alpha^{\overline{w}}$ *is denoted by* **Choice**$_\alpha^{\overline{w}}(h_v)$*. Observe that this implies that, for all v, v' in $\overline{w}$, $vR_\alpha^{\overline{w}}v'$ iff $h_v \in$* **Choice**$_\alpha^{\overline{w}}(h_{v'})$*. Similarly, observe that, for every $S \in \texttt{State}_\alpha^{\overline{w}}$, $S^T \in$* **State**$_\alpha^{\overline{w}}$*, and that, for every $U \in$* **State**$_\alpha^{\overline{w}}$*, there exists $V \in \texttt{State}_\alpha^{\overline{w}}$ such that $U = V^T$.*

- *For $\alpha \in Ags$, $\sim_\alpha$ is a relation on $I(M_W \times H_W)$ defined as follows:*

$$\begin{aligned} \sim_\alpha \quad := \quad & \{(\langle W, h_v\rangle, \langle W, h_{v'}\rangle); v, v' \in W\} \cup \\ & \left\{\left(\langle \overline{w}, h_v\rangle, \left\langle \overline{w'}, h_{v'}\right\rangle\right); w, w' \in W \text{ and } v \approx_\alpha v'\right\} \cup \\ & \{(\langle z, h_z\rangle, \langle z, h_z\rangle); z \in W\}. \end{aligned}$$

  This definition entails that $\sim_\alpha$ is an equivalence relation for every $\alpha \in Ags$ and that, for all $w, w' \in W$ and $L \in \texttt{Choice}_\alpha^{\overline{w}}$, $v \in [\![L]\!]_\alpha^{\overline{w'}}$ iff $h_v \in \left[L^T\right]_\alpha^{\overline{w'}}$.

- **Value** : $H_W \to \mathbb{R}$ *is a function defined as follows: for $h_v \in H_W$,* **Value**$(h_v) = \texttt{Value}(v)$*. Endowed with this function, we define the subjective dominance orderings $\preceq_s$ and $\prec_s$ according to Definition 4.20.*

Proposition C.44. *Let $\mathcal{F}$ be a Kripke-*eaus*-frame. Then $\mathcal{F}^T$ is indeed an* eaubt*-frame.*

Proof. It amounts to showing that $\sqsubset$ is a strict partial order that satisfies no backward branching, that **Choice** is a function that satisfies frame conditions $(\texttt{NC})$ and $(\texttt{IA})$, that $\{\sim_\alpha\}_{\alpha \in Ags}$ is such that $\sim_\alpha$ is an equivalence relation for every $\alpha \in Ags$ and frame conditions $(\texttt{OAC})$ and $(\texttt{Unif} - \texttt{H})$ are met, and that **Value** is well defined:

- As mentioned in Definition C.43, it is straightforward to show that $\sqsubset$ is a strict partial order that satisfies no backward branching. It is also clear from Definition C.43 that $\sim_\alpha$ is an equivalence relation for every $\alpha \in Ags$, and that **Value** is well defined.

- As for $(\texttt{NC})$, it is vacuously satisfied at moment W. It is satisfied in moments of the form $\overline{w}$ (with $w \in W$), since two different histories never intersect in a moment later than $\overline{w}$. Finally, it is also satisfied in moments of the form v such that $v \in W$ (since there are no moments above v).

- For ($\mathtt{IA}$), we reason by cases: (a) at moment W, ($\mathtt{IA}$) is validated straightforwardly, since $\mathbf{Choice}(\alpha, W) = \{H_W\}$ for every $\alpha \in Ags$; (b) for a moment of the form $\overline{w}$ (with $w \in W$), let s be a function that assigns to each agent α a member of $\mathbf{Choice}^{\overline{w}}_{\alpha} = \left\{C^T_\alpha ; C_\alpha \in \mathtt{Choice}^{\overline{w}}_{\alpha}\right\}$; let $s' : Ags \to \bigcup_{\alpha \in Ags} \mathtt{Choice}^{\overline{w}}_{\alpha}$ be a function such that $s'(\alpha) = C_\alpha$ iff $s(\alpha) = C^T_\alpha$; since $\mathcal{M}$ satisfies $(\mathtt{IA})_{\mathtt{K}}$, then $\bigcap_{\alpha \in Ags} s'(\alpha) \neq \emptyset$; take $v \in \bigcap_{\alpha \in Ags} s'(\alpha)$; then $v \in C_\alpha$ for every $\alpha \in Ags$; this implies that $h_v \in C^T_\alpha$ for every $\alpha \in Ags$, so that $\bigcap_{\alpha \in Ags} s(\alpha) \neq \emptyset$; (c) at moments of the form v such that $v \in W$, if s is a function that assigns to each α a member of $\mathbf{Choice}(v, \alpha)$, then s must be constant and $\bigcap_{\alpha \in Ags} s(\alpha) = \{h_v\}$.

- For ($\mathtt{OAC}$), take $\alpha \in Ags$. Again we reason by cases: (a) for indices based on moment W, ($\mathtt{OAC}$) is met straightforwardly, since $\sim_\alpha$ is defined so that $\langle W, h_v\rangle \sim_\alpha \langle W, h_{v'}\rangle$ for every pair of histories $h_v, h_{v'}$ in H_W; (b) for indices of the form $\langle \overline{w}, h_v\rangle$ (with $w \in W$ and $v \in \overline{w}$), we want to show that $\langle \overline{w}, h_v\rangle \sim_\alpha \langle \overline{w}, h_u\rangle$ for every $h_u \in \mathbf{Choice}^{\overline{w}}_{\alpha}(h_v)$; therefore, take $h_u \in \mathbf{Choice}^{\overline{w}}_{\alpha}(h_v)$, which implies that $u \in \mathtt{Choice}^{\overline{w}}_{\alpha}(v)$; since $\mathcal{M}$ satisfies $(\mathtt{OAC})_{\mathtt{K}}$, $v \approx_\alpha u$, which in turn yields that $\langle \overline{w}, h_v\rangle \sim_\alpha \langle \overline{w}, h_u\rangle$, by definition of $\sim_\alpha$; (c) for indices based on moments of the form v such that $v \in W$, ($\mathtt{OAC}$) is met straightforwardly, since for all $h_v \in H_W$ the choice-cell in $\mathbf{Choice}(\alpha, v)$ to which h_v belongs is just $\{h_v\}$.

- For ($\mathtt{Unif-H}$), take $\alpha \in Ags$. Again we reason by cases: (a) for indices based on moment W, ($\mathtt{Unif-H}$) is met straightforwardly, since $\langle W, h_v\rangle \sim_\alpha \langle W, h_{v'}\rangle$ for every $v, v' \in W$. (b) for indices of the form $\langle \overline{w}, h_v\rangle$ (with $w \in W$ and $v \in \overline{w}$), assume that $\langle \overline{w}, h_v\rangle \sim_\alpha \left\langle \overline{w'}, h_{v'}\right\rangle$; this means that $v \approx_\alpha v'$; take $h_z \in H_{\overline{w}}$ (which implies that $z \in \overline{w}$); we want to show that there exists $h \in H_{\overline{w'}}$ such that $\langle \overline{w}, h_z\rangle \sim_\alpha \left\langle \overline{w'}, h\right\rangle$; now, condition $(\mathtt{Unif-H})_{\mathtt{K}}$ for $\mathcal{M}$ gives that there exists $z' \in \overline{w'}$ such that $z \approx_\alpha z'$, which, by definition of $\sim_\alpha$, means that $\langle \overline{w}, h_z\rangle \sim_\alpha \left\langle \overline{w'}, h_{z'}\right\rangle$; since $z' \in \overline{w'}$ iff $h_z \in H_{\overline{w'}}$, we have shown what we wanted; (c) for indices based on moments of the form v such that $v \in W$, $\langle v, h_v\rangle$ is $\sim_\alpha$-related only to itself, so ($\mathtt{Unif-H}$) is met straightforwardly.

□

Let $\mathcal{M}$ be a Kripke-*eaus*-model with valuation function $\mathcal{V}$. The frame upon which $\mathcal{M}$ is based has an associated *eaubt*-frame. If to the tuple of this *eaubt*-frame one adds a valuation function $\mathcal{V}^T$ such that $\mathcal{V}^T(p) = \{\langle \overline{w}, h_w\rangle ; w \in \mathcal{V}(p)\}$, the resulting model is called the *eaubt*-model associated with $\mathcal{M}$.

Lemma C.45. *Let $\mathcal{M}$ be a Kripke-*eaus*-model, and let $\mathcal{M}^T$ be its associated* eaubt*-model. For all $\alpha \in Ags$, $w \in W$, and $L, N \in \mathtt{Choice}^{\overline{w}}_{\alpha}$, the following conditions hold:*

(a) $L \preceq_s N$ *iff* $L^T \preceq_s N^T$ *and* $L \prec_s N$ *iff* $L^T \prec_s N^T$.

(b) $L \in \mathtt{SOptimal}^{\overline{w}}_{\alpha}$ *iff* $L^T \in \mathbf{SOptimal}^{\overline{w}}_{\alpha}$.

Proof. (a) We prove this point only for the strict orderings, since this proof includes the arguments needed to show that the statement also holds for $\preceq_s$. ($\Rightarrow$) Assume that $L \prec_s N$. Let w' be such that $\overline{w} \sim_\alpha \overline{w'}$ (which implies that $\overline{w} \approx_\alpha \overline{w'}$),[36] and take $U \in \mathbf{State}^{\overline{w'}}_\alpha$. We know that $U = V^T$ for some $V \in \mathtt{State}^{\overline{w'}}_\alpha$. Our assumption implies that $\llbracket L \rrbracket^{\overline{w'}}_\alpha \cap V \leq \llbracket N \rrbracket^{\overline{w'}}_\alpha \cap V$. Recall that, for all $v \in W$, $\mathbf{Value}(h_v) = \mathtt{Value}(v)$ and that, for all $v_1, v_2 \in W$ such that $v_1 \in \llbracket L \rrbracket^{\overline{w'}}_\alpha \cap V$ and $v_2 \in \llbracket N \rrbracket^{\overline{w'}}_\alpha \cap V$, $h_{v_1} \in \left[L^T\right]^{\overline{w'}}_\alpha \cap V^T$ and $h_{v_2} \in \left[N^T\right]^{\overline{w'}}_\alpha \cap V^T$. Thus, the fact that $\llbracket L \rrbracket^{\overline{w'}}_\alpha \cap V \leq \llbracket N \rrbracket^{\overline{w'}}_\alpha \cap V$ implies that $\left[L^T\right]^{\overline{w'}}_\alpha \cap U \leq [N]^{\overline{w'}}_\alpha \cap U$. Now, our assumption also yields that there exist $w_* \in W^{\Lambda_S}$ and $S_0 \in \mathtt{State}^{\overline{w_*}}_\alpha$ such that $\overline{w} \approx_\alpha \overline{w_*}$ (which implies that $\overline{w} \sim_\alpha \overline{w_*}$) and such that $\llbracket N \rrbracket^{\overline{w_*}}_\alpha \cap S_0 \not\leq \llbracket L \rrbracket^{\overline{w_*}}_\alpha \cap S_0$. This implies that $\left[N^T\right]^{\overline{w_*}}_\alpha \cap S_0^T \not\leq \left[L^T\right]^{\overline{w_*}}_\alpha \cap S_0^T$, so that indeed $L^T \prec_s N^T$.

($\Leftarrow$) Assume that $L^T \prec_s N^T$. Let w' be such that $\overline{w} \approx_\alpha \overline{w'}$, and take $S \in \mathtt{State}^{\overline{w'}}_\alpha$. Our assumption implies that $\left[L^T\right]^{\overline{w'}}_\alpha \cap S^T \leq \left[N^T\right]^{\overline{w'}}_\alpha \cap S^T$. Thus, an argument similar to the one used for the left-to-right direction renders that $\llbracket L \rrbracket^{\overline{w'}}_\alpha \cap S \leq \llbracket N \rrbracket^{\overline{w'}}_\alpha \cap S$. Our assumption also implies that there exist $w_* \in W^{\Lambda_S}$ and $U_0 \in \mathbf{State}^{\overline{w_*}}_\alpha$ such that $\overline{w} \sim_\alpha \overline{w_*}$ (which implies that $\overline{w} \approx_\alpha \overline{w_*}$) and such that $\left[N^T\right]^{\overline{w_*}}_\alpha \cap U_0 \not\leq \left[L^T\right]^{\overline{w_*}}_\alpha \cap U_0$. But $U_0 = V_0^T$ for some $V_0 \in \mathtt{State}^{\overline{w_*}}_\alpha$. Thus, $\llbracket N \rrbracket^{\overline{w_*}}_\alpha \cap V_0 \not\leq \llbracket L \rrbracket^{\overline{w_*}}_\alpha \cap V_0$, so that indeed $L \prec_s N$.

(b) Straightforward, using point a above.

□

Proposition C.46 (Truth-preserving correspondence). *Let $\mathcal{M}$ be a Kripke-*eaus-*model, and let $\mathcal{M}^T$ be its associated* eaubt-*model. For all φ of $\mathcal{L}_S$ and $w \in W$, $\mathcal{M}, w \models \varphi$ iff $\mathcal{M}^T, \langle \overline{w}, h_w \rangle \models \varphi$.*

Proof. We proceed by induction on the complexity of φ. For the base case, take a propositional letter p and $w \in W$. Then $\mathcal{M}, w \models p$ iff $w \in \mathcal{V}(p)$ iff $\langle \overline{w}, h_w \rangle \in \mathcal{V}^T(p)$ iff $\mathcal{M}^T, \langle \overline{w}, h_w \rangle \models p$. The cases of Boolean connectives are standard, so let us deal with those of the modal operators. Take $w \in W$ and $\alpha \in Ags$.

[36] Recall that, for *eaubt*-models, I write $m \sim_\alpha m'$ if there exist $h \in H_m$ and $h' \in H_{m'}$ such that $\langle m, h \rangle \sim_\alpha \langle m', h' \rangle$.

- ("$\Box$") $\mathcal{M}, w \models \Box\varphi$ iff $\mathcal{M}, v \models \varphi$ for every $v \in \overline{w}$, which—by induction hypothesis—happens iff $\mathcal{M}^T, \langle \overline{v}, h_v \rangle \models \varphi$ for every $h_v \in H_{\overline{w}}$ (since $h_v \in H_{\overline{w}}$ iff $v \in \overline{w}$), which happens iff $\mathcal{M}^T, \langle \overline{w}, h_w \rangle \models \Box\varphi$.

- ("$[\alpha]$") $\mathcal{M}, w \models [\alpha]\varphi$ iff $\mathcal{M}, v \models \varphi$ for every $v \in W$ such that $wR_\alpha^{\overline{w}}v$, which—by induction hypothesis—happens iff $\mathcal{M}^T, \langle \overline{w}, h_v \rangle \models \varphi$ for every $h_v \in \mathbf{Choice}_\alpha^{\overline{w}}(h_w)$ (since $h_v \in \mathbf{Choice}_\alpha^{\overline{w}}(h_w)$ iff $wR_\alpha^{\overline{w}}v$), which in turn happens iff $\mathcal{M}^T, \langle \overline{w}, h_w \rangle \models [\alpha]\varphi$.

- ("K_α") $\mathcal{M}, w \models K_\alpha\varphi$ iff $\mathcal{M}, v \models \varphi$ for every $v \in W$ such that $w \approx_\alpha v$, which—by induction hypothesis—occurs iff $\mathcal{M}^T, \langle \overline{v}, h_v \rangle \models \varphi$ for every $h_v \in H$ such that $\langle \overline{w}, h_w \rangle \sim_\alpha \langle \overline{v}, h_v \rangle$ (since $\langle \overline{w}, h_w \rangle \sim_\alpha \langle \overline{v}, h_v \rangle$ iff $w \approx_\alpha v$), which happens iff $\mathcal{M}^T, \langle \overline{w}, h_w \rangle \models K_\alpha\varphi$.

- ("$\odot_\alpha^{\mathcal{S}}$") ($\Rightarrow$) Assume that $\mathcal{M}, w \models \odot_\alpha^{\mathcal{S}}\varphi$. Let $N^T \in \mathbf{Choice}_\alpha^{\overline{w}}$ be such that $[N]_\alpha^{\overline{w'}} \nsubseteq \left|\varphi\right|^{\overline{w'}}$ (for some w' such that $\overline{w} \sim_\alpha \overline{w'}$). The induction hypothesis implies that $[\![N]\!]_\alpha^{\overline{w'}} \nsubseteq \left|\varphi\right|$ (since $v \in [\![N]\!]_\alpha^{\overline{w'}}$ iff $h_v \in \left[N^T\right]_\alpha^{\overline{w'}}$). Our assumption then entails that there exists $L_1 \in \texttt{Choice}_\alpha^{\overline{w}}$ s. t. $N \prec_s L_1$ and, if $L = L_1$ or $L_1 \preceq_s L$, then $[\![L]\!]_\alpha^{\overline{w''}} \subseteq \left|\varphi\right|$ for every w'' such that $\overline{w} \approx_\alpha \overline{w''}$. We claim that L_1^T is the choice-cell at moment $\overline{w}$ that witnesses to the fact that $\mathcal{M}^T, \langle \overline{w}, h_w \rangle \models \odot_\alpha^{\mathcal{S}}\varphi$. To show this claim, first notice that Lemma C.45 a renders that $N^T \prec_s L_1^T$. Secondly, the fact that $[\![L_1]\!]_\alpha^{\overline{w''}} \subseteq \left|\varphi\right|$ for every w'' such that $\overline{w} \approx_\alpha \overline{w''}$ implies, with the induction hypothesis, that $\left[L_1^T\right]_\alpha^{\overline{w''}} \subseteq \left|\varphi\right|^{\overline{w'}}$ for every w'' such that $\overline{w} \sim_\alpha \overline{w''}$. Finally, let $L^T \in \mathbf{Choice}_\alpha^{\overline{w}}$ be such that $L_1^T \preceq_s L^T$. By Lemma C.45 a, $L_1 \preceq_s L$. Let w'' be such that $\overline{w} \sim_\alpha \overline{w''}$ (which implies that $\overline{w} \approx_\alpha \overline{w''}$). We know that $[\![L]\!]_\alpha^{\overline{w''}} \subseteq \left|\varphi\right|$, so the induction hypothesis implies that $\left[L^T\right]_\alpha^{\overline{w''}} \subseteq \left|\varphi\right|^{\overline{w''}}$. With this we have shown that our claim is true, so that $\mathcal{M}^T, \langle \overline{w}, h_w \rangle \models \odot_\alpha^{\mathcal{S}}\varphi$.

  ($\Leftarrow$) Assume that $\mathcal{M}^T, \langle \overline{w}, h_w \rangle \models \odot_\alpha^{\mathcal{S}}\varphi$. Let $N \in \texttt{Choice}_\alpha^{\overline{w}}$ be such that $[\![N]\!]_\alpha^{\overline{w'}} \nsubseteq \left|\varphi\right|$ (for some w' such that $w \approx_\alpha w'$). The induction hypothesis implies that $\left[N^T\right]_\alpha^{\overline{w'}} \nsubseteq \left|\varphi\right|^{\overline{w'}}$. Our assumption then entails that there exists $L_1^T \in \mathbf{Choice}_\alpha^{\overline{w}}$ such that $N^T \prec_s L_1^T$ and, if $L^T = L_1^T$ or $L_1^T \preceq_s L^T$, then $\left[L^T\right]_\alpha^{\overline{w''}} \subseteq \left|\varphi\right|^{\overline{w''}}$ for every w'' such that $\overline{w} \sim_\alpha \overline{w''}$. Here, we claim that L_1 witnesses to the fact that $\mathcal{M}, w \models \odot_\alpha^{\mathcal{S}}\varphi$. To show this, first notice that Lemma C.45 a renders that $N \prec_s L_1$. Secondly, the fact that $\left[L_1^T\right]_\alpha^{\overline{w''}} \subseteq \left|\varphi\right|^{\overline{w''}}$ for every w'' such that $\overline{w} \sim_\alpha \overline{w''}$ implies, with the induction hypothesis, that $[\![L_1]\!]_\alpha^{\overline{w''}} \subseteq \left|\varphi\right|$ for every

w'' such that $\overline{w} \approx_\alpha \overline{w''}$. Finally, let $L \in \mathtt{Choice}^{\overline{w}}_\alpha$ be such that $L_1 \preceq_s L$. By Lemma C.45 a, $L_1^T \preceq_s L^T$. Let w'' be such that $\overline{w} \approx_\alpha \overline{w''}$ (which implies that $\overline{w} \sim_\alpha \overline{w''}$). We know that $\left[L^T\right]_\alpha^{\overline{w''}} \subseteq \left|\varphi\right|^{\overline{w''}}$, so the induction hypothesis gives that $[\![L]\!]_\alpha^{\overline{w''}} \subseteq \left|\varphi\right|$. With this, we have shown that our claim is true, so that $\mathcal{M}, w \models \odot^S_\alpha \varphi$.

□

Proposition C.46 implies that to prove completeness of Λ_S with respect to *eaubt*-models all we need to do is prove completeness with respect to Kripke-*eaus*-models. Therefore, let us prove completeness with respect to Kripke-*eaus*-models, via the well-known technique of canonical models.

C.3.2 Canonical Kripke-*Eaus*-Structure

We show that the proof system Λ_S is complete with respect to the class of Kripke-*eaus*-models. For a Λ_S-consistent formula φ, we build a canonical structure that satisfies φ.

Definition C.47 (Canonical Structure). *The tuple*

$$\mathcal{M} = \left\langle W^{\Lambda_S}, R_\Box, \mathtt{Choice}, \{\approx_\alpha\}_{\alpha \in Ags}, \mathtt{Value}, \mathcal{V} \right\rangle$$

is called a canonical structure for Λ_S iff

- $W^{\Lambda_S} = \{w; w \text{ is a } \Lambda_S\text{-MCS}\}$. *$R_\Box$ is a relation on W^{Λ_S} defined by the rule: $wR_\Box v$ iff $\Box\varphi \in w \Rightarrow \varphi \in v$ for every φ of $\mathcal{L}_S$. For $w \in W^{\Lambda_S}$, the set $\left\{v \in W^{\Lambda_S}; wR_\Box v\right\}$ is denoted by $\overline{w}$.*

- $\mathtt{Choice}$ *is a function that assigns to each α and $\overline{w}$ a subset $\mathtt{Choice}^{\overline{w}}_\alpha$ of $2^{\overline{w}}$, defined as follows: let $R^{\overline{w}}_\alpha$ be a relation on $\overline{w}$ such that $wR^{\overline{w}}_\alpha v$ iff $[\alpha]\varphi \in w \Rightarrow \varphi \in v$ for every φ of $\mathcal{L}_S$; if $\mathtt{Choice}^{\overline{w}}_\alpha(v) := \left\{u \in \overline{w}; vR^{\overline{w}}_\alpha u\right\}$, then $\mathtt{Choice}^{\overline{w}}_\alpha := \left\{\mathtt{Choice}^{\overline{w}}_\alpha(v); v \in \overline{w}\right\}$.*

- *For $\alpha \in Ags$, $\approx_\alpha$ is a relation on W^{Λ_S} given by the rule: $w \approx_\alpha v$ iff $K_\alpha\varphi \in w \Rightarrow \varphi \in v$ for every φ of $\mathcal{L}_S$.*

- $\mathtt{Value}$ *is a function on W^{Λ_S} defined as follows: for $\alpha \in Ags$ and $w \in W^{\Lambda_S}$, we first define $\Gamma^w_\alpha := \left\{K_\alpha\varphi; \odot^S_\alpha\varphi \in w\right\}$ and $\Gamma^w := \bigcup_{\alpha \in Ags} \Gamma^w_\alpha$; then the function* $\mathtt{Value}$ *is given by the rule:*

$$\mathtt{Value}(w) \;=\; \begin{cases} 1 \text{ \textit{iff} } \Gamma^w \subseteq w \\ 0 \text{ \textit{otherwise.}} \end{cases}$$

- *Recall that P is the set of propositions in $\mathcal{L}_S$. Then $\mathcal{V} : P \to 2^{W^{\Lambda_S}}$ is the canonical valuation, defined so that $w \in \mathcal{V}(p)$ iff $p \in w$.*

Proposition C.48. *The canonical structure $\mathcal{M}$ for Λ_S is a Kripke-*eaus*-model.*

Proof. We want to show that the tuple $\left\langle W^{\Lambda_S}, R_\Box, \mathtt{Choice}, \{\approx_\alpha\}_{\alpha \in Ags}, \mathtt{Value} \right\rangle$ is a Kripke-*eaus*-frame, which amounts to showing that the tuple satisfies the items in the definition of Kripke-*eaus*-frames (Definition C.41).

- Since Λ_S includes the **S5** axioms for $\Box$, $R_\Box$ is an equivalence relation.

- Since Λ_S includes the **S5** schemata for $[\alpha]$, $R_\alpha^{\overline{w}}$ is an equivalence relation for all $\alpha \in Ags$ and $w \in W^{\Lambda_S}$. Moreover, since Λ_S includes schema (*SET*), $R_\alpha^{\overline{w}} \subseteq \overline{w} \times \overline{w}$ for every $w \in W^{\Lambda_S}$. Thus, $\mathtt{Choice}$ indeed assigns to each α and $\overline{w}$ a partition of $\overline{w}$. The fact that frame condition $(\mathtt{IA})_{\mathtt{K}}$ is satisfied is shown exactly as in Proposition A.15 (p. 118).

- Since the proof system Λ_S includes the **S5** schemata for K_α, $\approx_\alpha$ is an equivalence relation for every $\alpha \in Ags$. We verify that $\mathcal{M}$ satisfies conditions $(\mathtt{OAC})_{\mathtt{K}}$ and $(\mathtt{Unif-H})_{\mathtt{K}}$. For $(\mathtt{OAC})_{\mathtt{K}}$, take $w \in W^{\Lambda_S}$, $v \in \overline{w}$, and $\alpha \in Ags$. Take $u \in \mathtt{Choice}_\alpha^{\overline{w_*}}(v)$. This means that $vR_\alpha u$. Schema (*OAC*) and closure of v under *Modus Ponens* then implies that $[\alpha]\varphi \in v$. Since $vR_\alpha^{\overline{w}}u$, this implies that $\varphi \in v$. With this, we have shown that the fact that $K_\alpha\varphi \in v$ implies that $\varphi \in u$, which means that $v \approx_\alpha u$.

  $(\mathtt{Unif-H})_{\mathtt{K}}$ is shown to hold exactly as in Proposition A.15 (p. 120).

- $\mathtt{Value}$ is a well-defined function with range in $\mathbb{R}$.

□

Lemma C.49 (Existence for non-deontic operators)**.** *Let $\mathcal{M}$ be the canonical Kripke-*eaus*-model for Λ_S. For every $w \in W^{\Lambda_S}$ and every φ of $\mathcal{L}_{KO}$, the following items hold:*

1. *$\Box\varphi \in w$ iff $\varphi \in v$ for every $v \in \overline{w}$.*

2. *$[\alpha]\varphi \in w$ iff $\varphi \in v$ for every $v \in \overline{w}$ such that $wR_\alpha^{\overline{w}}v$.*

3. *$K_\alpha\varphi \in w$ iff $\varphi \in v$ for every $v \in W^{\Lambda_S}$ such that $w \approx_\alpha v$.*

Proof. The proof is the same as the one included for Lemma A.16 (p. 121). □

Observation C.50. *For $w \in W^{\Lambda_S}$ and $\alpha \in Ags$, let $\alpha_k[w]$ denote the set $\left\{v \in W^{\Lambda_S}; \Diamond K_\alpha\varphi \in w \Rightarrow \Diamond K_\alpha\varphi \in v\right\}$. Then the set $\alpha_k[w]$ is the same as the set $\left\{v \in W^{\Lambda_S}; wR_\Box \circ \approx_\alpha v\right\}$.*[37]

Proof. For the $\subseteq$ inclusion, take $v \in \alpha_k[w]$. Let us show that the set $u' := \{\varphi; K_\alpha\varphi \in w\} \cup \{\theta; \Box\theta \in v\}$ is consistent. Suppose for a contradiction that it is not consistent. In virtue of arguments analogous to the one used in Proposition A.15's proof of item $(\texttt{Unif}-\texttt{H})_\texttt{K}$ (p. 120), we know that $\{\varphi; K_\alpha\varphi \in w\}$ and $\{\theta; \Box\theta \in v\}$ are consistent. Thus, there must exist sets $\{\varphi_1, \dots, \varphi_n\}$ and $\{\theta_1, \dots, \theta_m\}$ of formulas of $\mathcal{L}_\mathsf{S}$ such that (a) $K_\alpha\varphi_i \in w$ for every $1 \leq i \leq n$, (b) $\Box\theta_i \in v$ for every $1 \leq i \leq m$, and (c) $\vdash_{\Lambda_S} (\varphi_1 \wedge \cdots \wedge \varphi_n) \wedge (\theta_1 \wedge \cdots \wedge \theta_m) \to \bot$. Let $\varphi = \varphi_1 \wedge \cdots \wedge \varphi_n$, and let $\theta = \theta_1 \wedge \cdots \wedge \theta_m$. On the one hand, since K_α and $\Box$ distribute over conjunction, it is the case that $\vdash_{\Lambda_S} K_\alpha\varphi \leftrightarrow K_\alpha\varphi_1 \wedge \cdots \wedge K_\alpha\varphi_n$ and that $\vdash_{\Lambda_S} \Box\theta \leftrightarrow \Box\theta_1 \wedge \cdots \wedge \Box\theta_m$. On the other hand, Λ_S-theorem (c) implies that $\vdash_{\Lambda_S} \varphi \to \neg\theta$ and thus that (d) $\vdash_{\Lambda_S} \Diamond K_\alpha\varphi \to \Diamond K_\alpha\neg\theta$. Observe that, since w is a Λ_S-MCS closed under conjunction and logical equivalence, $K_\alpha\varphi \in w$. Similarly, since v is closed under conjunction and logical equivalence, ($\star$) $\Box\theta \in v$. The fact that $K_\alpha\varphi \in w$, with schema (T) for $\Box$ and closure of w under *Modus Ponens*, implies that $\Diamond K_\alpha\varphi \in w$. The fact that $v \in \alpha_k[w]$ then implies that $\Diamond K_\alpha\varphi \in v$, so that Λ_S-theorem (d) and closure of v under *Modus Ponens* imply that $\Diamond K_\alpha\neg\theta \in v$. Schema $(Unif - H)$ and closure of v under *Modus Ponens* then imply that $K_\alpha\Diamond\neg\theta \in v$. This last fact implies, by schema (T) for K_α and closure of v under *Modus Ponens*, that $\Diamond\neg\varphi \in v$, contradicting ($\star$). Therefore, u' is in fact consistent. Let u be the Λ_S-MCS that includes u', which exists in virtue of Lindenbaum's Lemma (Blackburn et al., 2002, Chapter 4, p. 199). By construction, $u \in \overline{v}$ and $w \approx_\alpha u$. Thus, $wR_\Box \circ \approx_\alpha v$.

For the $\supseteq$ inclusion, let $v \in W^{\Lambda_S}$ be such that $wR_\Box \circ \approx_\alpha v$. Assume that $\Diamond K_\alpha\varphi \in w$. This assumption implies that $K_\alpha\Box\Diamond K_\alpha\varphi \in w$ according to the following argument: schema (4) for K_α and the fact that $\vdash_{\Lambda_S} (p \to q) \to (\Diamond p \to \Diamond q)$ imply that $\vdash_{\Lambda_S} \Diamond K_\alpha\varphi \to \Diamond K_\alpha K_\alpha\varphi$. Schema (5) for $\Box$ and transitivity of implication then imply that $\vdash_{\Lambda_S} \Diamond K_\alpha\varphi \to \Box\Diamond K_\alpha K_\alpha\varphi$.[38] Schema $(Unif - H)$, the fact that $\vdash_{\Lambda_S} (p \to q) \to (\Box p \to \Box q)$, and transitivity of implication then imply that $\vdash_{\Lambda_S} \Diamond K_\alpha\varphi \to \Box K_\alpha\Diamond K_\alpha\varphi$. The fact that $\vdash_{\Lambda_S} K_\alpha\Box p \leftrightarrow \Box K_\alpha p$—which was shown to be true in Observation 4.31 b—and transitivity of implication then imply that $\vdash_{\Lambda_S} \Diamond K_\alpha\varphi \to K_\alpha\Box\Diamond K_\alpha\varphi$. Therefore, closure of w under *Modus Ponens* implies that

[37] Recall from Chapter 3 (Footnote 36) that, for relations R, S on a given set, I write $R \circ S$ to denote the composition of R and S, such that $x(R \circ S)y$ iff there exists z in the relevant set such that xSz and zRy.

[38] I use the term 'transitivity of implication' to refer to the fact that $\models_{\Lambda_S} ((p \to q) \wedge (q \to r)) \to (p \to r)$, which is clear from propositional logic.

$K_\alpha \Box \Diamond K_\alpha \varphi \in w$. Let $u \in W^{\Lambda_S}$ be such that $w \approx_\alpha u$ and such that $u \in \overline{v}$. The facts that $K_\alpha \Box \Diamond K_\alpha \varphi \in w$ and that $w \approx_\alpha u$ imply that $\Box \Diamond K_\alpha \varphi \in u$. Therefore, the fact that $u \in \overline{v}$ implies that $\Diamond K_\alpha \varphi \in v$. This shows that $v \in \alpha_k[w]$. □

Lemma C.51 (Existence for subjective ought-to-do's). *For all $\alpha \in Ags$ and $w \in W^{\Lambda_S}$, the following points hold:*

(a) For every φ of $\mathcal{L}_\mathsf{S}$, $\odot^S_\alpha \varphi \in w$ iff $\varphi \in v$ for every $v \in \alpha_k[w]$ such that $\Gamma^v_\alpha \subseteq v$.

(b) For all w' such that $\overline{w} \approx_\alpha \overline{w'}$ and all $v \in \left[\!\left[\mathtt{Choice}^{\overline{w}}_\alpha(w) \right]\!\right]^{\overline{w'}}_\alpha$, $\Gamma^w_\alpha \subseteq w$ iff $\Gamma^v_\alpha \subseteq v$.

(c) $\Gamma^w_\alpha \subseteq w$ iff $\mathtt{Choice}^{\overline{w}}_\alpha(w) \in \mathtt{SOptimal}^{\overline{w}}_\alpha$.

(d) For every $L \in \mathtt{Choice}^{\overline{w}}_\alpha - \mathtt{SOptimal}^{\overline{w}}_\alpha$, there exists $L' \in \mathtt{SOptimal}^{\overline{w}}_\alpha$ such that $L \prec_s L'$. Thus, $\mathcal{M}, w \models \odot^S_\alpha \varphi$ iff $[\![L]\!]^{\overline{w'}}_\alpha \subseteq \left|\varphi\right|$ for every $L \in \mathtt{SOptimal}^{\overline{w}}_\alpha$ and every w' such that $w \approx w'$.

Proof. (a) Before the proof of this point, we need to show that two preliminary claims are true:

Claim 1: for every $v \in \alpha_k[w]$, $\Gamma^v_\alpha = \Gamma^w_\alpha$. *Proof of claim:* take $v \in \alpha_k[w]$. For the $\supseteq$ inclusion, assume that $K_\alpha \varphi \in \Gamma^w_\alpha$. This means that $\odot^S_\alpha \varphi \in w$. By schema (*s.Cl*) and closure of w under *Modus Ponens*, the fact that $\odot^S_\alpha \varphi \in w$ implies that $K_\alpha \Box \odot^S_\alpha \varphi \in w$. Since $v \in \alpha_k[w]$, Observation C.50 implies that $\odot^S_\alpha \varphi \in v$. This then implies that $K_\alpha \varphi \in \Gamma^v_\alpha$. The other inclusion is analogous, since the fact that $v \in \alpha_k[w]$ implies that $w \in \alpha_k[v]$ according to the following argument: $(\mathtt{Unif-H})_\mathtt{K}$ implies that $R_\Box \circ \approx_\alpha = \approx_\alpha \circ R_\Box$; thus, $\approx_\alpha \circ R_\Box$ is an equivalence relation, so that Observation C.50 implies that $v \in \alpha_k[w]$ iff $w \in \alpha_k[v]$ *(end of proof of claim)*.

Claim 2: for any pair of agents $\alpha, \beta \in Ags$ and any formula φ of $\mathcal{L}_\mathsf{S}$, $\odot^S_\alpha \varphi \to \neg \odot^S_\beta \neg \varphi \in w$ for every Λ_S-MCS w. We refer to this property as *consistency of subjective ought-to-do*. *Proof of claim*: this claim is a direct consequence from the fact that, for every $\alpha, \beta \in Ags$ and every φ a formula of $\mathcal{L}_\mathsf{S}$, $\vdash_{\Lambda_S} \odot^S_\alpha \varphi \to \neg \odot^S_\beta \neg \varphi$. In the case where $\alpha \neq \beta$, this comes from the fact that, for every $\alpha, \beta \in Ags$, $\vdash_{\Lambda_S} \odot^S_\alpha \varphi \wedge \odot^S_\beta \neg \varphi \to \bot$—which can be seen by applying schemata (*s.Oic*), (*OAC*), (*IA*), and (*T*) for $[\alpha]$ and $[\beta]$. In the case where $\alpha = \beta$, this can be seen by noticing that $\odot^S_\alpha$ distributes over conjunction, so that (*s.Oic*), schema (*T*) for K_α, and the fact that $\vdash_{\Lambda_S} (p \to q) \to (\Diamond p \to \Diamond q)$ yield the required Λ_S-theorem *(end of proof of claim)*.

Now, we proceed with the proof of the main statement. Let $\odot^S_\alpha \varphi$ be a formula of $\mathcal{L}_\mathsf{S}$.

($\Rightarrow$) Assume that $\odot^{S}_{\alpha}\varphi \in w$. Schema (*s.Cl*) and closure of w under *Modus Ponens* implies that $K_{\alpha}\Box \odot^{S}_{\alpha} \varphi \in w$. The fact that $v \in \alpha_k[w]$ and Observation C.50 yield that $\odot^{S}_{\alpha}\varphi \in v$. The assumption that $\Gamma^{v}_{\alpha} \subseteq v$ then entails that $K_{\alpha}\varphi \in v$. Schema (T) for K_{α} and closure of v under *Modus Ponens* imply that $\varphi \in v$.

($\Leftarrow$) We work by contraposition. Suppose that $\odot^{S}_{\alpha}\varphi \notin w$. We first show that Γ^{w}_{α} is consistent. Suppose that Γ^{w}_{α} is not consistent. Then there is a set $\{\varphi_1, \ldots, \varphi_n\}$ of formulas of $\mathcal{L}_{\mathsf{S}}$ such that (a) $\odot^{S}_{\alpha}\varphi_i \in w$ for every $1 \leq i \leq n$, and (b) $\vdash_{\Lambda_S} K_{\alpha}\varphi_1 \wedge \cdots \wedge K_{\alpha}\varphi_n \rightarrow \bot$. This last Λ_S-theorem implies that $\vdash_{\Lambda_S} K_{\alpha}\varphi_1 \wedge \cdots \wedge K_{\alpha}\varphi_{n-1} \rightarrow \neg K_{\alpha}\varphi_n$. By Necessitation and schema (K) for $\odot^{S}_{\alpha}$, as well as its distributivity over conjunction, it is then the case that

$$\vdash_{\Lambda_S} \odot^{S}_{\alpha}(K_{\alpha}\varphi_1) \wedge \cdots \wedge \odot^{S}_{\alpha}(K_{\alpha}\varphi_{n-1}) \rightarrow \odot^{S}_{\alpha}(\neg K_{\alpha}\varphi_n). \tag{4.1}$$

The fact that $\odot^{S}_{\alpha}\varphi_i \in w$ for every $1 \leq i \leq n-1$ implies, by schema ($A6$) and closure of w under *Modus Ponens*, that $\odot^{S}_{\alpha}(K_{\alpha}\varphi_i) \in w$ for every $1 \leq i \leq n-1$. Closure of w under conjunction then implies that $\left(\bigwedge_{1\leq i\leq n-1} \odot^{S}_{\alpha}(K_{\alpha}\varphi_i)\right) \in w$. With Λ_S-theorem (4.1) and closure of w under *Modus Ponens*, this implies that $\odot^{S}_{\alpha}(\neg K_{\alpha}\varphi_n) \in w$. However, schema ($A6$) and closure of w under *Modus Ponens* yield that the fact that $\odot^{S}_{\alpha}\varphi_n \in w$ implies that $\odot^{S}_{\alpha}(K_{\alpha}\varphi_n) \in w$. Thus, we both have that $\odot^{S}_{\alpha}(\neg K_{\alpha}\varphi_n) \in w$ and that $\odot^{S}_{\alpha}(K_{\alpha}\varphi_n) \in w$, which is a contradiction, according to *consistency of subjective ought-to-do* (**Claim 2**). Next, we show that $\Gamma^{w}_{\alpha} \cup \{\Diamond K_{\alpha}\theta; \Diamond K_{\alpha}\theta \in w\}$ is also consistent. To prove this, suppose that it is not consistent. Since Γ^{w}_{α} and $\{\Diamond K_{\alpha}\theta; \Diamond K_{\alpha}\theta \in w\}$ are consistent, there must exist sets $\{\varphi_1, \ldots, \varphi_n\}$ and $\{\theta_1, \ldots, \theta_m\}$ of formulas of $\mathcal{L}_{\mathsf{S}}$ such that (a) $\odot^{S}_{\alpha}\varphi_i \in w$ for every $1 \leq i \leq n$, (b) $\Diamond K_{\alpha}\theta_i \in w$ for every $1 \leq i \leq m$, and (c) $\vdash_{\Lambda_S} (K_{\alpha}\varphi_1 \wedge \cdots \wedge K_{\alpha}\varphi_n) \wedge (\Diamond K_{\alpha}\theta_1 \wedge \cdots \wedge \Diamond K_{\alpha}\theta_m) \rightarrow \bot$. By Necessitation and schema (K) for $\odot^{S}_{\alpha}$, as well as its distributivity over conjunction, it is then the case that

$$\vdash_{\Lambda_S} \left(\bigwedge_{1\leq i\leq n} \odot^{S}_{\alpha}(K_{\alpha}\varphi_i)\right) \wedge \left(\bigwedge_{1\leq i\leq m} \odot^{S}_{\alpha}(\Diamond K_{\alpha}\theta_i)\right) \rightarrow \odot^{S}_{\alpha}\bot. \tag{4.2}$$

The fact that $\odot^{S}_{\alpha}\varphi_i \in w$ for every $1 \leq i \leq n$ implies, with schema ($A6$) and closure of w under *Modus Ponens*, that $\odot^{S}_{\alpha}(K_{\alpha}\varphi_i) \in w$ for every $1 \leq i \leq n$. Closure of w under conjunction then implies that ($\star$) $\left(\bigwedge_{1\leq i\leq n} \odot^{S}_{\alpha}(K_{\alpha}\varphi_i)\right) \in w$. As mentioned in Observation C.50, the fact that $\Diamond K_{\alpha}\theta_i \in w$ for every $1 \leq i \leq m$ implies that $K_{\alpha}\Box\Diamond K_{\alpha}\theta_i \in w$ for every $1 \leq i \leq m$. Therefore, closure of w under *Modus Ponens* implies that $K_{\alpha}\Box\Diamond K_{\alpha}\theta_i \in w$ for every $1 \leq i \leq m$. Schema (*SuN*) and closure of w under *Modus Ponens* then imply that $\odot^{S}_{\alpha}(\Diamond K_{\alpha}\theta_i) \in w$ for every $1 \leq i \leq m$, so that closure of w under conjunction implies that ($\star\star$)

$\left(\bigwedge_{1\leq i\leq m} \odot^{S}_{\alpha} (\Diamond K_{\alpha}\theta_i)\right) \in w$. Hence, we have struck a contradiction, because ($\star$) and ($\star\star$) imply that the antecedent in Λ_S-theorem (4.2) lies in w, so that closure of w under *Modus Ponens* would imply that $\odot^{S}_{\alpha}\bot \in w$. Therefore, $\Gamma^{w}_{\alpha} \cup \{\Diamond K_{\alpha}\theta; \Diamond K_{\alpha}\theta \in w\}$ is consistent.

Finally, we show that $\Gamma^{w}_{\alpha} \cup \{\Diamond K_{\alpha}\theta; \Diamond K_{\alpha}\theta \in w\} \cup \{\neg\varphi\}$ is also consistent. First, recall that our main assumption is that $\odot^{S}_{\alpha}\varphi \notin w$. Now suppose that $\Gamma^{w}_{\alpha} \cup \{\Diamond K_{\alpha}\theta; \Diamond K_{\alpha}\theta \in w\} \cup \{\neg\varphi\}$ is not consistent. Since $\Gamma^{w}_{\alpha} \cup \{\Diamond K_{\alpha}\theta; \Diamond K_{\alpha}\theta \in w\}$ is consistent, there must exist sets $\{\varphi_1, \ldots, \varphi_n\}$ and $\{\theta_1, \ldots, \theta_m\}$ of formulas of $\mathcal{L}_{\mathsf{S}}$ such that (a) $\odot^{S}_{\alpha}\varphi_i \in w$ for every $1 \leq i \leq n$, (b) $\Diamond K_{\alpha}\theta_i \in w$ for every $1 \leq i \leq m$, and (c) $\vdash_{\Lambda_S} (K_{\alpha}\varphi_1 \wedge \cdots \wedge K_{\alpha}\varphi_n) \wedge (\Diamond K_{\alpha}\theta_1 \wedge \cdots \wedge \Diamond K_{\alpha}\theta_m) \wedge \neg\varphi \to \bot$. Now, this Λ_S-theorem implies that $\vdash_{\Lambda_S} (K_{\alpha}\varphi_1 \wedge \cdots \wedge K_{\alpha}\varphi_n) \wedge (\Diamond K_{\alpha}\theta_1 \wedge \cdots \wedge \Diamond K_{\alpha}\theta_m) \to \varphi$. By Necessitation and schema (K) for $\odot^{S}_{\alpha}$, as well as its distributivity over conjunction, it is then the case that

$$\vdash_{\Lambda_S} \left(\bigwedge_{1\leq i\leq n} \odot^{S}_{\alpha} (K_{\alpha}\varphi_i)\right) \wedge \left(\bigwedge_{1\leq i\leq m} \odot^{S}_{\alpha} (\Diamond K_{\alpha}\theta_i)\right) \to \odot^{S}_{\alpha}\varphi. \tag{4.3}$$

With similar arguments to the ones used before, the fact that $\odot^{S}_{\alpha}\varphi_i \in w$ for every $1 \leq i \leq n$ implies that $\odot^{S}_{\alpha} (K_{\alpha}\varphi_i) \in w$ for every $1 \leq i \leq n$, and the fact that $\Diamond K_{\alpha}\theta_i \in w$ for every $1 \leq i \leq m$ implies that $\odot^{S}_{\alpha} (\Diamond K_{\alpha}\theta_i) \in w$ for every $1 \leq i \leq m$. Closure of w under conjunction then implies that $\left(\bigwedge_{1\leq i\leq n} \odot^{S}_{\alpha} (K_{\alpha}\varphi_i)\right) \in w$ and that $\left(\bigwedge_{1\leq i\leq m} \odot^{S}_{\alpha} (\Diamond K_{\alpha}\theta_i)\right) \in w$. Therefore, closure of w under conjunction implies that the antecedent in Λ_S-theorem (4.3) lies in w, so that closure of w under *Modus Ponens* entails that $\odot^{S}_{\alpha}\varphi \in w$. But this contradicts the assumption that $\odot^{S}_{\alpha}\varphi \notin w$. Therefore, $\Gamma^{w}_{\alpha} \cup \{\Diamond K_{\alpha}\theta; \Diamond K_{\alpha}\theta \in w\} \cup \{\neg\varphi\}$ is in fact consistent.

Let u be the Λ_S-MCS that includes $\Gamma^{w}_{\alpha} \cup \{\Diamond K_{\alpha}\theta; \Diamond K_{\alpha}\theta \in w\} \cup \{\neg\varphi\}$, which exists in virtue of Lindenbaum's Lemma (Blackburn et al., 2002, Chapter 4, p. 199). By construction, $u \in \alpha_k[w]$. This last fact implies, by **Claim 1**, that $\Gamma^{u}_{\alpha} = \Gamma^{w}_{\alpha}$, so that our construction guarantees that $\Gamma^{u}_{\alpha} \subseteq u$ and that $\varphi \notin u$. In this way, we have shown that assuming that $\odot^{S}_{\alpha} \notin w$ implies the existence of $u \in \alpha_k[w]$ such that $\Gamma^{u}_{\alpha} \subseteq u$ and such that $\varphi \notin u$.

(b) Let $w' \in W^{\Lambda_S}$ be such that $\overline{w} \approx_{\alpha} \overline{w'}$, and take $v \in \left[\!\left[\mathtt{Choice}^{\overline{w}}_{\alpha}(w)\right]\!\right]^{\overline{w'}}_{\alpha}$. ($\Rightarrow$) Assume that $\Gamma^{w}_{\alpha} \subseteq w$. First, observe that taking v in $\left[\!\left[\mathtt{Choice}^{\overline{w}}_{\alpha}(w)\right]\!\right]^{\overline{w'}}_{\alpha}$ implies that $w \approx_{\alpha} v$, according to the following argument: by definition, if v lies in $[\![\mathtt{Choice}_{\alpha}(w)]\!]^{\overline{w}}_{\alpha}$, then there exists $v' \in \mathtt{Choice}^{\overline{w}}_{\alpha}(w)$ such that $v' \approx_{\alpha} v$; condition $(\mathtt{OAC})_{\mathtt{K}}$ renders that $w \approx_{\alpha} v'$, so transitivity of $\approx_{\alpha}$ implies that $w \approx_{\alpha} v$. Now, take $K_{\alpha}\varphi \in \Gamma^{w}_{\alpha}$.

By assumption, $K_\alpha\varphi \in w$. Schema (4) for K_α and closure of w under *Modus Ponens* implies that $K_\alpha K_\alpha\varphi \in w$ as well. Since $w \approx_\alpha v$, this in turn yields that $K_\alpha\varphi \in v$. Therefore, we have shown that $\Gamma^w_\alpha \subseteq v$, but then point a's **Claim 1** implies that $\Gamma^v_\alpha \subseteq v$, since the fact that $w \approx_\alpha v$ implies that $v \in \alpha_k[w]$ (as shown in the proof of **Claim 1**). ($\Leftarrow$) Analogous.

(c) Before the proof of this point, we need to show that a preliminary claim is true:

Claim 3: for every $w \in W^{\Lambda_S}$, the set $\{u \in \overline{w}; \texttt{Value}(u) = 1\}$ is not empty. *Proof of claim*: take $w \in W^{\Lambda_S}$. First, we show that $\Gamma^w = \bigcup_{\alpha\in Ags}\Gamma^w_\alpha$ is consistent. To prove this, suppose that it is not consistent. Then for each $\alpha \in Ags$ there exists a set $\{\varphi_{1_\alpha}, \ldots, \varphi_{n_\alpha}\}$ of formulas of $\mathcal{L}_S$ such that $\bigwedge\limits_{1_\alpha\leq i_\alpha\leq n_\alpha} \odot^S_\alpha\varphi_{i_\alpha} \in w$, and such that (i) $\vdash_{\Lambda_S} \bigwedge\limits_{\alpha\in Ags}\left(\bigwedge\limits_{1_\alpha\leq i_\alpha\leq n_\alpha} K_\alpha\varphi_{i_\alpha}\right) \to \bot$. For $\alpha \in Ags$, take $\varphi_\alpha = \varphi_{1_\alpha}\wedge\cdots\wedge\varphi_{n_\alpha}$. Since K_α distributes over conjunction, Λ_S-theorem (i) implies that (ii) $\vdash_{\Lambda_S} \bigwedge\limits_{\alpha\in Ags} K_\alpha\varphi_\alpha \to \bot$. Since w is closed under conjunction, the fact that $\bigwedge\limits_{1_\alpha\leq i_\alpha\leq n_\alpha} \odot^S_\alpha\varphi_{i_\alpha} \in w$ for every $\alpha \in Ags$ implies that $\bigwedge\limits_{\alpha\in Ags}\left(\bigwedge\limits_{1_\alpha\leq i_\alpha\leq n_\alpha} \odot^S_\alpha\varphi_{i_\alpha}\right) \in w$. Since w is closed under logical equivalence, this implies that $\bigwedge\limits_{\alpha\in Ags} \odot^S_\alpha\varphi_\alpha \in w$. Observe, then, that schema (*s.Oic*) yields that (iii) $\vdash_{\Lambda_S} \bigwedge\limits_{\alpha\in Ags} \odot^S_\alpha\varphi_\alpha \to \bigwedge\limits_{\alpha\in Ags} \Diamond K_\alpha\varphi_\alpha$. Now, for all $\alpha \in Ags$, schema (4) for K_α and schema (*OAC*) yield that $\vdash_{\Lambda_S} K_\alpha\rho \to [\alpha]K_\alpha\rho$, and thus that $\vdash_{\Lambda_S} \Diamond K_\alpha\rho \to \Diamond[\alpha]K_\alpha\rho$. Therefore, Λ_S-theorem (iii) and transitivity of implication imply that (iv) $\vdash_{\Lambda_S} \bigwedge\limits_{\alpha\in Ags} \odot^S_\alpha\varphi_\alpha \to \bigwedge\limits_{\alpha\in Ags} \Diamond[\alpha]K_\alpha\varphi_\alpha$. Now, observe that schema (*IA*) implies that (v) $\vdash_{\Lambda_S} \bigwedge\limits_{\alpha\in Ags} \Diamond[\alpha]K_\alpha\varphi_\alpha \to \Diamond\left(\bigwedge\limits_{\alpha\in Ags} [\alpha]K_\alpha\varphi_\alpha\right)$, and schemata (*T*) for $[\alpha]$ (for all $\alpha \in Ags$) imply that (vi) $\vdash_{\Lambda_S} \Diamond\left(\bigwedge\limits_{\alpha\in Ags} [\alpha]K_\alpha\varphi_\alpha\right) \to \bigwedge\limits_{\alpha\in Ags} K_\alpha\varphi_\alpha$. Therefore, by transitivity of implication, Λ_S-theorems (iv), (v), (vi), and (ii) imply that $\vdash_{\Lambda_S} \bigwedge\limits_{\alpha\in Ags} \odot^S_\alpha\varphi_\alpha \to \Diamond\bot$. But this is a contradiction, since $\bigwedge\limits_{\alpha\in Ags} \odot^S_\alpha\varphi_\alpha \in w$, and w is a Λ_S-MCS. Therefore, Γ^w is consistent. Next, we show that the union $\Gamma^w \cup \{\theta; \Box\theta \in w\}$ is also consistent. To prove this, suppose that it is not consistent. Since Γ^w and $\{\theta; \Box\theta \in w\}$ are consistent, then there exist sets $\{\varphi_{1_\alpha}, \ldots, \varphi_{n_\alpha}\}$ (for each $\alpha \in Ags$) and $\{\theta_1, \ldots, \theta_m\}$ of formulas of $\mathcal{L}_S$ such that (a) $\bigwedge_{1_\alpha \leq i_\alpha \leq n_\alpha} \odot^S_\alpha \varphi_{i_\alpha} \in w$ for every $\alpha \in Ags$, (b) $\Box\theta_i \in w$ for every $1 \leq i \leq m$, and (c) $\vdash_{\Lambda_S} \bigwedge_{\alpha \in Ags} \left(\bigwedge_{1_\alpha \leq i_\alpha \leq n_\alpha} K_\alpha \varphi_{i_\alpha} \right) \wedge (\theta_1 \wedge \cdots \wedge \theta_m) \to \bot$. Let $\theta = \theta_1 \wedge \cdots \wedge \theta_m$. Since $\Box$ distributes over conjunction, one has that $\vdash_{\Lambda_S} \Box\theta \leftrightarrow \Box\theta_1 \wedge \cdots \wedge \Box\theta_m$, where it is important to mention that closure of w under conjunction and logical equivalence implies that $\Box\theta \in w$. With this equivalence, Λ_S-theorem (c) implies that (d) $\vdash_{\Lambda_S} \bigwedge_{\alpha \in Ags} \left(\bigwedge_{1_\alpha \leq i_\alpha \leq n_\alpha} K_\alpha \varphi_{i_\alpha} \right) \to \neg\theta$. Again, for $\alpha \in Ags$, take $\varphi_\alpha = \varphi_{1_\alpha} \wedge \cdots \wedge \varphi_{n_\alpha}$. By an argument that is analogous to the one used to show that Γ^w is consistent, Λ_S-theorem (d) implies that (e) $\vdash_{\Lambda_S} \bigwedge_{\alpha \in Ags} \odot^S_\alpha \varphi_\alpha \to \Diamond\neg\theta$. Again, since w is closed under conjunction and logical equivalence, the fact that $\bigwedge_{1_\alpha \leq i_\alpha \leq n_\alpha} \odot^S_\alpha \varphi_{i_\alpha} \in w$ for every $\alpha \in Ags$ implies that $\bigwedge_{\alpha \in Ags} \odot^S_\alpha \varphi_\alpha \in w$. By closure of w under *Modus Ponens*, Λ_S-theorem (e) entails that $\Diamond\neg\theta \in w$, but this is a contradiction, since we had seen that $\Box\theta \in w$. Therefore, $\Gamma^w \cup \{\theta; \Box\theta \in w\}$ is consistent.

Let u be the Λ_S-MCS that includes $\Gamma^w \cup \{\theta; \Box\theta \in w\}$, which exists in virtue of Lindenbaum's Lemma (Blackburn et al., 2002, Chapter 4, p. 199). By construction, $u \in \overline{w}$ and $\Gamma^w \subseteq u$. Point a's **Claim 1** and the fact that $u \in \overline{w}$ yield that $\Gamma^u = \Gamma^w$. Therefore, $\Gamma^u \subseteq u$, so that $\mathtt{Value}(u) = 1$. In this way, we have shown that the set $\{u \in \overline{w}; \mathtt{Value}(u) = 1\}$ is not empty *(end of proof of claim)*.

For the proof of the main statement, first recall that $\mathtt{SOptimal}^{\overline{w}}_\alpha$ is defined as $\left\{L \in \mathtt{Choice}^{\overline{w}}_\alpha; \text{there is no } L' \in \mathtt{Choice}^{\overline{w}}_\alpha \text{ such that } L \prec_s L'\right\}$.

($\Rightarrow$) Assume that $\Gamma^w_\alpha \subseteq w$. We want to show that, for all $L \in \mathtt{Choice}^{\overline{w}}_\alpha$, $L \preceq_s \mathtt{Choice}^{\overline{w}}_\alpha(w)$, since this implies that $\mathtt{Choice}^{\overline{w}}_\alpha(w) \in \mathtt{SOptimal}^{\overline{w}}_\alpha$. Take $L \in \mathtt{Choice}^{\overline{w}}_\alpha$. To prove what we want, we show that, for all w' such that $\overline{w} \approx_\alpha \overline{w'}$, all $S \in \mathtt{State}^{\overline{w'}}_\alpha$, and all u, o such that $u \in [\![L]\!]^{\overline{w'}}_\alpha \cap S$ and $o \in \left[\!\left[\mathtt{Choice}^{\overline{w}}_\alpha(w)\right]\!\right]^{\overline{w'}}_\alpha \cap S$, $\mathtt{Value}(u) \leq \mathtt{Value}(o)$. For this, it suffices to show that, for all w' such that $\overline{w} \approx_\alpha \overline{w'}$ and all $S \in \mathtt{State}^{\overline{w'}}_\alpha$, if $\mathtt{Value}(u_*) = 1$ for some $u_* \in [\![L]\!]^{\overline{w'}}_\alpha \cap S$, then $\mathtt{Value}(o) = 1$ for every $o \in \left[\!\left[\mathtt{Choice}^{\overline{w}}_\alpha(w)\right]\!\right]^{\overline{w'}}_\alpha \cap S$. Therefore, let w' be such that $\overline{w} \approx_\alpha \overline{w'}$. Take $S \in \mathtt{State}^{\overline{w'}}_\alpha$, $o \in \left[\!\left[\mathtt{Choice}^{\overline{w}}_\alpha(w)\right]\!\right]^{\overline{w'}}_\alpha \cap S$, and $u_* \in [\![L]\!]^{\overline{w'}}_\alpha \cap S$ such that $\mathtt{Value}(u_*) = 1$. By point a's **Claim 1**, this last assumption means that $\Gamma^{u_*} = \Gamma^{w'} = \bigcup_{\alpha \in Ags} \Gamma^{w'}_\alpha \subseteq u_*$. We show that $\mathtt{Value}(o) = 1$. We first prove that, for every $v \in S$, $\bigcup_{\beta \in Ags - \{\alpha\}} \Gamma^v_\beta \subseteq v$: take $v \in S$; this means that $v \in \mathtt{Choice}^{\overline{w'}}_\beta(u_*)$ for every $\beta \in Ags - \{\alpha\}$; by $(\mathtt{OAC})_{\mathtt{K}}$, this entails that

$v \in \left\llbracket \mathtt{Choice}^{\overline{w'}}_{\beta}(u_*) \right\rrbracket^{\overline{w'}}_{\beta}$ for every $\beta \in Ags - \{\alpha\}$; thus, point b of the present lemma implies that $\Gamma^{v}_{\beta} \subseteq v$ for every $\beta \in Ags - \{\alpha\}$, so that $\bigcup_{\beta \in Ags-\{\alpha\}} \Gamma^{v}_{\beta} \subseteq v$. Now, since o was taken in $\left\llbracket \mathtt{Choice}^{\overline{w}}_{\alpha}(w) \right\rrbracket^{\overline{w'}}_{\alpha} \cap S \subseteq S$, then, on the one hand, the observation just made implies that $(\star)$ $\bigcup_{\beta \in Ags-\{\alpha\}} \Gamma^{o}_{\beta} \subseteq o$; on the other, the fact that $o \in \left\llbracket \mathtt{Choice}^{\overline{w}}_{\alpha}(w) \right\rrbracket^{\overline{w'}}_{\alpha} \cap S \subseteq \left\llbracket \mathtt{Choice}^{\overline{w}}_{\alpha}(w) \right\rrbracket^{\overline{w'}}_{\alpha}$ implies, by point b of the present lemma, that $(\star\star)$ $\Gamma^{o}_{\alpha} \subseteq o$. By $(\star)$ and $(\star\star)$, $\Gamma^{o} = \bigcup_{\alpha \in Ags} \Gamma^{o}_{\alpha} \subseteq o$, so that $\mathtt{Value}(o) = 1$, which is what we wanted to show.

$(\Leftarrow)$ We work by contraposition. Assume that $\Gamma^{w}_{\alpha} \not\subseteq w$. We want to show that there exists an action in $\mathtt{Choice}^{\overline{w}}_{\alpha}$ that dominates $\mathtt{Choice}^{\overline{w}}_{\alpha}(w)$ in the subjective ordering. By point b of the present lemma, our assumption implies that, for every $w' \in W^{\Lambda_s}$ such that $\overline{w} \approx_{\alpha} \overline{w'}$, $\Gamma^{o}_{\alpha} \not\subseteq o$ for every $o \in \left\llbracket \mathtt{Choice}^{\overline{w}}_{\alpha}(w) \right\rrbracket^{\overline{w'}}_{\alpha}$. This means that, for all $w' \in W^{\Lambda_s}$ such that $\overline{w} \approx_{\alpha} \overline{w'}$, $\mathtt{Value}(o) = 0$ for every $o \in \left\llbracket \mathtt{Choice}^{\overline{w}}_{\alpha}(w) \right\rrbracket^{\overline{w'}}_{\alpha}$. From **Claim 3** we know that the set $\{u \in \overline{w}; \mathtt{Value}(u) = 1\}$ is not empty. Take $u_* \in \{u \in \overline{w}; \mathtt{Value}(u) = 1\}$. We claim that $\mathtt{Choice}^{\overline{w}}_{\alpha}(u_*)$ is the action that we are looking for. To prove this claim, we check that two conditions hold: (1) for all $w' \in W^{\Lambda_s}$ such that $\overline{w} \approx_{\alpha} \overline{w'}$ and all $S \in \mathtt{State}^{\overline{w'}}_{\alpha}$, $\left\llbracket \mathtt{Choice}^{\overline{w}}_{\alpha}(w) \right\rrbracket^{\overline{w'}}_{\alpha} \cap S \leq \left\llbracket \mathtt{Choice}^{\overline{w}}_{\alpha}(u_*) \right\rrbracket^{\overline{w'}}_{\alpha} \cap S$, and (2) there exist $w_* \in W^{\Lambda_s}$ and $S_* \in \mathtt{State}^{\overline{w_*}}_{\alpha}$ such that $\overline{w} \approx_{\alpha} \overline{w_*}$ and such that $\left\llbracket \mathtt{Choice}^{\overline{w}}_{\alpha}(u_*) \right\rrbracket^{\overline{w_*}}_{\alpha} \cap S_* \nleq \left\llbracket \mathtt{Choice}^{\overline{w}}_{\alpha}(w) \right\rrbracket^{\overline{w_*}}_{\alpha} \cap S_*$. For (1), let $w' \in W^{\Lambda_s}$ be such that $\overline{w} \approx_{\alpha} \overline{w'}$, and take $S \in \mathtt{State}^{\overline{w'}}_{\alpha}$. Since $\mathtt{Value}(o) = 0$ for every $o \in \left\llbracket \mathtt{Choice}^{\overline{w}}_{\alpha}(w) \right\rrbracket^{\overline{w'}}_{\alpha}$, then $\left\llbracket \mathtt{Choice}^{\overline{w}}_{\alpha}(w) \right\rrbracket^{\overline{w'}}_{\alpha} \cap S \leq \left\llbracket \mathtt{Choice}^{\overline{w}}_{\alpha}(u_*) \right\rrbracket^{\overline{w'}}_{\alpha} \cap S$. For (2), let $S_{u_*} := \bigcap_{\beta \in Ags-\{\alpha\}} \mathtt{Choice}^{\overline{w}}_{\beta}(u_*)$. Observe that $\overline{w} \approx_{\alpha} \overline{w}$, that $S_{u_*} \in \mathtt{State}^{\overline{w}}_{\alpha}$, and that $u_* \in \left\llbracket \mathtt{Choice}^{\overline{w}}_{\alpha}(u_*) \right\rrbracket^{\overline{w}}_{\alpha} \cap S_{u_*}$. In turn, $(\mathtt{IA})_{\mathrm{K}}$ implies that $\mathtt{Choice}^{\overline{w}}_{\alpha}(w) \cap S_{u_*} \neq \emptyset$, so that $(\mathtt{OAC})_{\mathrm{K}}$ gives that $\left\llbracket \mathtt{Choice}^{\overline{w}}_{\alpha}(w) \right\rrbracket^{\overline{w}}_{\alpha} \cap S_{u_*} \neq \emptyset$. Since $\mathtt{Value}(o) = 0$ for every $o \in \left\llbracket \mathtt{Choice}^{\overline{w}}_{\alpha}(w) \right\rrbracket^{\overline{w}}_{\alpha} \cap S_{u_*}$, and since $\mathtt{Value}(u_*) = 1$, $\left\llbracket \mathtt{Choice}^{\overline{w}}_{\alpha}(u_*) \right\rrbracket^{\overline{w}}_{\alpha} \cap S_{u_*} \nleq \left\llbracket \mathtt{Choice}^{\overline{w}}_{\alpha}(w) \right\rrbracket^{\overline{w}}_{\alpha} \cap S_{u_*}$. Thus, results (1) and (2) render that $\mathtt{Choice}^{\overline{w}}_{\alpha}(w) \prec_s \mathtt{Choice}^{\overline{w}}_{\alpha}(u_*)$, and this implies that $\mathtt{Choice}^{\overline{w}}_{\alpha}(w) \notin \mathtt{SOptimal}^{\overline{w}}_{\alpha}$.

(d) Take $L \in \mathtt{Choice}^{\overline{w}}_{\alpha} - \mathtt{SOptimal}^{\overline{w}}_{\alpha}$. By **Claim 3**, there exists $u_* \in \overline{w}$ such that $\mathtt{Value}(u_*) = 1$. Point c of the present lemma implies that $\mathtt{Choice}^{\overline{w}}_{\alpha}(u_*) \in \mathtt{SOptimal}^{\overline{w}}_{\alpha}$. Observe that the fact that $L \in \mathtt{Choice}^{\overline{w}}_{\alpha} - \mathtt{SOptimal}^{\overline{w}}_{\alpha}$ implies, with points b and c of the present lemma, that $\mathtt{Value}(o) = 0$ for every w' such that $\overline{w} \approx_{\alpha} \overline{w'}$ and $o \in \llbracket L \rrbracket^{\overline{w'}}_{\alpha}$, according to the following argument: if

$L \in \mathtt{Choice}^{\overline{w}}_{\alpha} - \mathtt{SOptimal}^{\overline{w}}_{\alpha}$, then point c yields that $(\star)$ $\Gamma^{u}_{\alpha} \not\subseteq u$ for every $u \in L$; now, let $w' \in W^{\Lambda_S}$ be such that $\overline{w} \approx_{\alpha} \overline{w'}$; take $o \in [\![L]\!]^{\overline{w'}}_{\alpha}$, which means that there exists $u_o \in L$ such that $u_o \approx_{\alpha} o$; according to $(\star)$, $\Gamma^{u_o}_{\alpha} \not\subseteq u_o$; point b then implies that $\Gamma^{o}_{\alpha} \not\subseteq o$, so that $\mathtt{Value}(o) = 0$. Therefore, one has that (1) for all w' such that $\overline{w} \approx_{\alpha} \overline{w'}$ and all $S \in \mathtt{State}^{\overline{w'}}_{\alpha}$, $[\![L]\!]^{\overline{w'}}_{\alpha} \cap S \leq \left[\!\left[\mathtt{Choice}^{\overline{w}}_{\alpha}(u_*)\right]\!\right]^{\overline{w'}}_{\alpha} \cap S$. On the other hand, if S_* denotes the unique state in $\mathtt{State}^{\overline{w}}_{\alpha}$ such that $u_* \in S_*$, then it is clear that (2) $\left[\!\left[\mathtt{Choice}^{\overline{w}}_{\alpha}(u_*)\right]\!\right]^{\overline{w}}_{\alpha} \cap S_* \not\leq [\![L]\!]^{\overline{w}}_{\alpha} \cap S_*$. Our results (1) and (2) render that $L <_s \mathtt{Choice}^{\overline{w}}_{\alpha}(u_*)$.

□

Lemma C.52 (Truth Lemma). *Let $\mathcal{M}$ be the canonical Kripke-*eaus*-model for Λ_S. For all φ of $\mathcal{L}_S$ and $w \in W^{\Lambda_S}$, $\mathcal{M}, w \models \varphi$* iff *$\varphi \in w$.*

Proof. We proceed by induction on the complexity of φ. The cases of propositional letters and of Boolean connectives are standard. It remains to deal with the modal operators. For formulas involving $\Box$, $[\alpha]$, and K_{α}, both directions follow straightforwardly from Lemma C.49 (items 1, 2, and 3, respectively). As for the case of $\odot^{S}_{\alpha}$, we have the following arguments:

- ("$\odot^{S}_{\alpha}$") ($\Rightarrow$) We work by contraposition. Assume that $\odot^{S}_{\alpha}\varphi \notin w$. By Lemma C.51 a, there exists $v \in \alpha_k[w]$ such that $\Gamma^{v}_{\alpha} \subseteq v$ and such that $\varphi \notin v$. Let $o_v \in \overline{w}$ be such that $v \approx_{\alpha} o_v$—where we know that such an o_v exists because Observation C.50 and condition $(\mathtt{Unif-H})_{\mathtt{K}}$ guarantee that the fact that $v \in \alpha_k[w]$ implies that there exists o_v in $\overline{w}$ such that $v \approx_{\alpha} o_v$. Observe that the fact that $o_v \approx_{\alpha} v$ implies that $v \in \left[\!\left[\mathtt{Choice}^{\overline{w}}_{\alpha}(o_v)\right]\!\right]^{\overline{v}}_{\alpha}$. Now, by induction hypothesis, $\mathcal{M}, v \not\models \varphi$. By Lemma C.51 b, the facts that $\Gamma^{v}_{\alpha} \subseteq v$ and $v \approx_{\alpha} o_v$ imply that $\Gamma^{o_v}_{\alpha} \subseteq o_v$, so that Lemma C.51 c then implies that $\mathtt{Choice}^{\overline{w}}_{\alpha}(o_v) \in \mathtt{SOptimal}^{\overline{w}}_{\alpha}$. This means that there exists an action in $\mathtt{SOptimal}^{\overline{w}}_{\alpha}$—namely $\mathtt{Choice}^{\overline{w}}_{\alpha}(o_v)$—such that $v \in \left[\!\left[\mathtt{Choice}^{\overline{w}}_{\alpha}(o_v)\right]\!\right]^{\overline{v}}_{\alpha}$ and such that $\mathcal{M}, v \not\models \varphi$, which by Lemma C.51 d implies that $\mathcal{M}, w \not\models \odot^{S}_{\alpha}\varphi$.

  ($\Leftarrow$) Assume that $\odot^{S}_{\alpha}\varphi \in w$. Because of Lemma C.51 d, we want to show that $[\![L]\!]^{\overline{w'}}_{\alpha} \subseteq |\varphi|$ for every $L \in \mathtt{SOptimal}^{\overline{w}}_{\alpha}$ and $w' \in W^{\Lambda_S}$ such that $\overline{w} \approx_{\alpha} \overline{w'}$. Thus, take $L \in \mathtt{SOptimal}^{\overline{w}}_{\alpha}$, and let $w' \in W^{\Lambda_S}$ be such that $\overline{w} \approx_{\alpha} \overline{w'}$. By Lemma C.51 a, our assumption implies that $\varphi \in v$ for every $v \in \alpha_k[w]$ such that $\Gamma^{v}_{\alpha} \subseteq v$. By Lemma C.51 c and b, $\Gamma^{v'}_{\alpha} \subseteq v'$ for every $v' \in [\![L]\!]^{\overline{w'}}_{\alpha}$, where Observation C.50 implies that $[\![L]\!]^{\overline{w'}}_{\alpha} \subseteq \alpha_k[w]$. Thus, $\varphi \in v'$ for every $v' \in [\![L]\!]^{\overline{w'}}_{\alpha}$. By induction hypothesis, then, $\mathcal{M}, v' \models \varphi$ for every $v' \in [\![L]\!]^{\overline{w'}}_{\alpha}$, which means that $[\![L]\!]^{\overline{w'}}_{\alpha} \subseteq |\varphi|$. Thus, $\mathcal{M}, w \models \odot^{S}_{\alpha}\varphi$.

□

Proposition C.53 (Completeness w.r.t. Kripke-*eaus*-models)**.** *The proof system Λ_S is complete with respect to the class of Kripke-*eaus*-models.*

Proof. Let φ be a Λ_S-consistent formula of $\mathcal{L}_S$. Let w be the Λ_S-MCS including φ, which exists in virtue of Lindenbaum's Lemma (Blackburn et al., 2002, Chapter 4, p. 199). Then the canonical structure $\mathcal{M}$ for Λ_S is a Kripke-*eaus*-model such that $\mathcal{M}, w \models \varphi$, according to Lemma C.52 above. □

Proposition C.54 (Completeness w.r.t. *eaubt*-models)**.** *The proof system Λ_S is complete with respect to the class of* eaubt-*models.*

Proof. Let φ be a Λ_S-consistent formula of $\mathcal{L}_S$. Proposition C.53 implies that there exists a Kripke-*eaus*-model $\mathcal{M}$ and a world w in its domain such that $\mathcal{M}, w \models \varphi$. Proposition C.46 then ensures that the *eaubt*-model $\mathcal{M}^T$ associated with $\mathcal{M}$ is such that $\mathcal{M}^T, \langle \overline{w}, h_w \rangle \models \varphi$. □

Therefore, Proposition C.40 and Proposition C.54 imply that the following result, appearing in the main body of the chapter, has been shown:

Theorem 4.30 (Soundness & Completeness of Λ_S)**.** *The proof system Λ_S is sound and complete with respect to the class of* eaubt-*models.*

5

Agency, Knowledge, and Intentionality

'And yet it disturbs me to learn I have hurt someone unintentionally. I want all my hurts to be intentional.'

Margaret Atwood, *Cat's Eye*

5.1 Introduction

Suppose that you are a lawyer. You are part of the prosecution in a trial where the defendant is being accused of murder. The case is as follows: while driving her car, the defendant ran over and killed a traffic officer who was standing at a crossing walk. At the trial, the defense is seeking for a charge of *involuntary manslaughter*, and the prosecution contends that it was either *second-* or *first-degree murder*.[1] This

[1] According to *Wikipedia* (`https://en.wikipedia.org/wiki/Murder_in_United_States_law`), American law distinguishes the following degrees of murder, whose descriptions are included verbatim:

- *First-degree murder*: any intentional killing that is willful and premeditated with malice aforethought.
- *Second-degree murder*: any intentional killing that is not premeditated or planned.
- *Voluntary manslaughter*: sometimes called a crime of passion murder, it is any intentional killing that involves no prior intent to kill, and which was committed under such circumstances that would 'cause a reasonable person to become emotionally or mentally disturbed.' Both this and second-degree murder are committed on the spot under a spur-of-the-moment choice, but the two differ in the magnitude of the circumstances surrounding the crime. For example, a bar fight that results in death would ordinarily constitute second-degree murder. If that same bar fight stemmed from a discovery of infidelity, however, it may be voluntary manslaughter.
- *Involuntary manslaughter*: a killing that stems from a lack of intention to cause death but involving an intentional or negligent act leading to death. A drunk driving-related death is

means that the verdict revolves around the intentionality of the defendant. If the prosecuting team—to which you belong—is able to prove that the defendant had an intention to kill the officer, then the verdict would be either of second- or first-degree murder (according to whether the murder was either planned or unplanned). If the defense shows that the evidence does not support that there was an intention to kill—as would be the case if, for instance, the defendant was drunk while driving and had no real motive for killing the officer—then the verdict would be of only involuntary manslaughter.

This example shows that, as far as responsibility attribution in criminal law goes, intentionality is of the utmost importance. For many reasons, this importance has carried over to philosophy, giving rise to an ongoing debate as to the relation between intentionality and responsibility. This chapter is devoted to the incorporation into stit theory of agents' *intentions* and *intentional actions*, which are key components of responsibility according to the decomposition presented on p. 3. On the road to building my formal theory of responsibility, such an incorporation will help in characterizing the category of *motivational responsibility* (see the discussion on categories of responsibility in Chapter 1, p. 5).

To clarify, recall that an agent is motivationally responsible for a state of affairs iff the agent is the material author of such a state and the agent behaved knowingly and intentionally while bringing it about. Thus, to formalize motivational responsibility, a formalization of what it means to *intentionally* bring about a state of affairs is necessary. This chapter includes a stit-theoretic proposal for the latter formalization, where an agent's intentional actions are defined in terms of what the agent intends at a specific moment of acting. In other words, I define intentional actions in terms of *present-directed intentions* (Bratman, 1984; Broersen, 2011b; Lorini & Herzig, 2008).

The main goal of this chapter, then, is to provide an axiomatizable (stit-theoretic) logic to reason about the interplay between three essential components of responsibility: agency, knowledge, and intentionality. Thus, here I extend epistemic stit theory (*EST*) (see Chapter 2's Subsection 2.4.4) with modality $I_\alpha\varphi$, meant to express that at a given point in time agent α had a present-directed intention toward the realization of φ. The semantics for $I_\alpha\varphi$ is based on special topologies, each associated with an agent, that are added to *ebt*-frames (see Defini-

typically involuntary manslaughter. Note that the 'unintentional' element here refers to the lack of intent to bring about the death. If there is a presence of intention, it relates only to the intent to cause a violent act which brings about the death, but not an intention to bring about the death itself.

tion 2.27). For a given agent, the non-empty open sets of the associated topology are interpreted as present-directed intentions. I refer to the resulting formalism as *intentional epistemic stit theory* (*IEST*). An outline of the chapter is included below.

- Section 5.2 briefly reviews philosophy of intention's main ideas (and problems) around the notions of *intending* and *intentionally doing*, paying special attention to previous logic-based frameworks.
- Section 5.3 introduces my theory of intentionality (*IEST*). Since such a theory represents intentions with open sets in specific topologies, the pertinent definitions of General Topology are addressed. Examples designed to illustrate the basic aspects of *IEST* are also explored.
- Section 5.4 discusses *IEST*'s logic-based and metalogic properties. A Hilbert-style proof system for the logic is investigated, as well as its soundness & completeness results.
- Section 5.5 (the conclusion) discusses two possibilities for future work: first, the modelling of future-directed intentions using temporal stit theory; secondly, an extension of *IEST* with a probabilistic semantics of belief. Furthermore, an initial characterization of motivational responsibility is presented.

5.2 A Bit of Background: Philosophy of Intention

What do we talk about when we talk about intentions, intending, intentional action, and intentionally doing? Well, philosophy of intention has a lot to say about these concepts and about their interplay. For the sake of clarity as to succeeding sections' discussions (and terminology), in this section I provide some philosophical background on intentionality, as well as on previous logic-based approaches to modelling it.

In the opening lines of the current *SEP* entry for intention, Setiya (2018) wrote:

> Philosophical perplexity about intention begins with its appearance in three guises: intention for the future, as when I intend to complete this entry by the end of the month; the intention with which someone acts, as I am typing with the further intention of writing an introductory sentence; and intentional action, as in the fact that I am typing these words intentionally.

In the philosophical literature it is well-known that modelling intentionality is difficult, that it involves many interesting issues, and that no camp has the last word on what the best framework for analyzing the concept is. However, most authors agree with the quote above, and thus identify three main forms of intentionality:

1. *Future-directed intentions*: closely following the interpretation of Bratman (1984, 1987), I take future-directed intentions as elements that make up an agent's *plans*.[2] In other words, future-directed intentions are mental attitudes that an agent has towards possible states of affairs that lie in the future, that help in the coordinating of said agent's activities for bringing about those states of affairs. In the quote above, when the author mentions that he intends to complete his entry by the end of the month, the word 'intends' refers to future-directed intentions. Now, the literature also acknowledges the existence of *present-directed intentions*, referring to mental states that regard what agents intend to do now. Just as Bratman (1984), here I opt to include present-directed intentions in the category of future-directed intentions.[3]

2. *Intentional action*: following Broersen (2011b), who put forward a logic-based account of claims advanced by Anscombe (1963), I interpret intentional action as a mode of acting. Such a mode sets apart actions that are done with the purpose of bringing about some of the states of affairs that ensue from them, on the one hand, and actions that are done without any explicit goal of that kind, on the other. In the quote above, when the author mentions that he types words while writing his entry, and that he is doing so intentionally, he is referring to the intentional action of hitting a keyboard's keys.

3. *Intention-with-which*: following Davidson (1980), I interpret intention-with-which as a description of the primary reason that an agent has for acting in a specific way. In other words, intention-with-which refers to the motivation underlying a particular choice of action, what the agent seeks to bring about with such an action. In the quote above, when the author mentions that he types words toward the goal of writing an introductory sentence, then writing an introductory sentence is an intention-with-which he types.

[2] Bratman (1984, p. 379, emphasis in original) wrote that there is "ambiguity in talk about plans. Sometimes we are talking about *states* of the agent—states of *having* certain plans. Other times we are talking about an appropriate abstract structure—some sort of partial function from circumstances to actions, perhaps—that may be used to describe the planning-states of different people." According to Bratman, a more careful usage reserves the term 'plan' for the latter notion and 'having a plan' for the former. Here, just as Bratman, I refer to the state of 'having a plan.'

[3] Bratman (1984, p. 379) wrote: "[n]ote that even my present-directed intention to start my car is an intention to perform an action that continues somewhat into the future."

Intuitively speaking, these three guises are distinct but closely related to one another. The typical example that sets future-directed intentions apart from intentional action is that I might intend to start my car today and actually never come to do it. Thus, my intention did not translate into an action, let alone an instance of intentional action. The same example sets apart future-directed intentions from intention-with-which, since intending does not imply doing. In turn, the typical example that sets intention-with-which apart from intentional action is that I might be intentionally moving my hand and turning the key in the ignition, but the intention-with-which I am performing these actions is to start my car.[4]

One of the main problems in philosophy of intention, then, has been to find unity in these three senses of intentionality. According to Setiya (2018), finding unity matters for questions in philosophy of mind, but also in ethics, in epistemology, and for questions about the nature of practical reason.[5]

The literature includes notable attempts to solve this unity-problem. In such attempts, a few camps have formed. For instance, Davidson (1963, 1978) famously put forward that, while one can explain intention-with-which in terms of intentional action, one cannot reduce future-directed intentions to intentional action. Roughly speaking, Davidson (1963) described agents as having primary reasons for acting the way they do, and the relation between a given action and those primary reasons is what renders the action as either intentional or unintentional. As for intentions-with-which, he considered them as mere descriptions of the primary reasons. As for future-directed intentions, however, Davidson (1978) considered them to be a whole different matter, since they admit the following two facets: (1) as mentioned above, one can intend something without taking any action toward achieving it, and (2) it is still the case that future-directed intentions can be present in intentional action (as when I intend to start my car and then just go and do it). Davidson's work led many authors to seek an explanation of unity by treating future-directed intentions as primitive and then defining intentional action in terms of future-directed intentions (see, for instance, Aune, 1977; Bratman, 1984, 1987; Searle, 1983). All these authors belong to the same camp in the question of unity, then.

Other authors, however, showed resistance to the view that intentional action should be defined on the basis of future-directed intentions. The seminal work

[4] Anscombe (1963) wrote that, since it is implausible to say that the word 'intention' is equivocal, then the fact that it has different senses tells us that "we are pretty much in the dark about the character of the concept which it represents" (Anscombe, 1963, p. 1).

[5] Practical reason refers to the capacity for deciding, through reflection, what one is to do. Given a set of alternatives for action, *none of which has yet been executed*, practical reason is employed by an agent to settle on what the agent ought to do, or what action is best (Wallace, 2020). Observe that the study of practical reason is closely related to the study of ought-to-do presented in Chapter 4.

of Anscombe (1963) gave birth to a long-standing tradition in philosophy of action, where intentional action is primitive and where both intention-with-which and future-directed intentions are defined in terms of intentional action (see, for instance, Falvey, 2000; Moran & Stone, 2009; Thompson, 2008). According to Setiya (2018), the simplest version of this approach emphasizes the 'openness' of the progressive tense: if someone *is doing* φ, then this does not imply that they will succeed in doing φ, or even that they are well on their way to achieving φ. One can therefore *identify* future-directed intentions with intentional action, because the latter can refer to an action that has just begun and will not necessarily bring the effects intended by the future-directed intention. If I intend to play basketball today, I am already on the way to doing so, but it is possible that something—such as writing this chapter in my thesis—gets in the way. Thompson (2008) further argued that, even if the openness of the progressive tense is not invoked, future-directed intentions are actually processes 'in progress' toward the intentional completion of an act. Thus, the only difference between future-directed intentions and intentional actions is that expressions of the latter kind imply some measure of success.

Although prevalent, the quest for unity is not the only problem that philosophers of intention deal with. Additionally, there are big challenges in describing the relation between intentions, knowledge, belief, desires, volition, and evaluative judgement. For an overview of how authors have approached these challenges, the reader is once again referred to Setiya (2018).

To correctly position the following section's proposal, I should also review some of the logic-based studies of intentionality. Among them, it is important to mention the *BDI* logics (Cohen & Levesque, 1990; Herzig & Longin, 2004; Meyer et al., 1999; Rao & Georgeff, 1991; Shoham, 1993; Wooldridge, 2000). These frameworks are prominent in the literature on multi-agent systems (MAS), and they rely on the assumption that intelligent agents' choices are influenced by mental constructs such as beliefs, desires, and intentions. Actually, the acronym 'BDI' stands precisely for *beliefs-desires-intentions*. Following Bratman's (1984) philosophical theory of intentionality, *BDI* logics typically focus on future-directed intentions. These intentions are seen as mental states of agents, that are constituents of more complex plans, and they are modelled as particular sets either in branching-time frames or in domains of possible worlds where each world represents a course of events. The language of these logics, then, includes a modality of the form $\alpha\ int : \varphi$, meant to express that agent α intends to realize φ.

For instance, Cohen and Levesque's (1990) seminal paper introduced a first-order modal logic of beliefs and goals. Deeply inspired by dynamic logic, Cohen and Levesque's models include a set of agents, a set of possible worlds, a set of action types, and functions underlying the basic modalities for agents' beliefs and for agents' goals.[6] Within this framework, an agent's future-directed intention—of performing a given action type—is characterized as a specific combination of the modality for belief and the modality for goals. As argued by Broersen (2011b), although the approach enabled the authors to reason about several important properties of action—like pre- and post- conditions in the context of action-type composition—it excluded a thorough exploration of intentional action.[7]

Even if they do not exactly fall into the category of *BDI* logics, I would like to address two other logic-based frameworks that share with *BDI* Bratman's main intuition of prioritizing future-directed intentions over both intentional action and intention-with-which. These frameworks were respectively given by Konolige and Pollack (1993) and by Duijf (2018, Chapter 4), and they are sound attempts to model future-directed intentions using *neighborhood semantics* that yield non-normal modal logics of intentionality (see Montague, 1970; Scott, 1970). In the first case, the use of neighborhood semantics came from an interest in solving a philosophical problem (for the formalization of intention) known as the *side-effect problem* (Bratman, 1987; Broersen, 2011b; Cohen & Levesque, 1990). Since the side-effect problem revolves around the questions of whether intentions should be closed under belief, knowledge, or logical consequence,[8] Konolige and Pollack (1993) opted for a non-normal modal logic of intention where agents do not intend all the logical consequences of whatever they intend. In the second case, Duijf (2018, Chapter 4) introduced a notion of *admissibility of actions* with respect to an agent's given intention, such that an action is admissible with respect to the intention if no other action is strictly better suited to fulfilling that intention. Since it is generally impossible for an agent to perform an action that is both

[6] The syntax of Cohen and Levesque's logic includes action types in the object language, so their logic is indeed closely related to dynamic logic.

[7] Most of the *BDI* frameworks mentioned here only focus on the semantic aspect of the developed logics. Their proponents were not particularly concerned with issues of axiomatization. A notable exception is the propositional fragment of Cohen and Levesque's seminal logic given by Herzig and Longin (2004). Indeed, the authors of the latter work presented a sound and complete proof system for their logic of beliefs, goals, and intentions.

[8] According to Bratman (1987), it is clear that an agent who intends to perform an action usually does not intend all the consequences of that action, or even all the consequences that the agent anticipates. Some of the consequences are indeed goals of the agent, while others are 'side effects' that the agent is not committed to. The typical example supporting this view involves an agent intending to go to get his tooth filled at the dentist. Being uninformed about anesthetics, the agent believes that the process of having his tooth filled will necessarily cause him much pain. Although the agent intends to ask the dentist to fill his tooth, and, believing what he does, he is willing to put up with pain, the agent would surely deny that he intends to be in pain (see Cohen & Levesque, 1990, p. 218).

admissible with respect to its intention and admissible with respect to all the logical consequences of such an intention, Duijf also favored a non-normal modal logic based on neighborhood semantics.[9]

In contrast to the logic-based approaches mentioned above, which prioritized future-directed intentions, Broersen (2011b) set out to model the notion of intentional action instead. By means of extending *EST* with a combined modality for intentional action—$I_\alpha[\alpha]\varphi$—Broersen explored the relation between *intentionally seeing to it that* φ and *knowingly doing* φ. His aim—which is very similar to my own—was to make a start on the analysis of responsibility in the context of the modes of *mens rea*. In his formalization, the partition of an agent's available choices at some moment is independent of the partition based on the equivalence relation underlying modality $I_\alpha\varphi$, so that intentional action is characterized with $I_\alpha[\alpha]\varphi$, and unintentional action is characterized with $[\alpha]\varphi \wedge \neg I_\alpha[\alpha]\varphi$. Admittedly, Broersen (2011b) also gave an account—however implicit—of mere intending with $I_\alpha\varphi$, but he did not explore it thoroughly and rather focused on $I_\alpha[\alpha]\varphi$.[10]

Along the same lines, Lorini and Herzig (2008) modelled intentional action with a formalism that is technically similar to Herzig and Longin's (2004) propositional fragment of Cohen and Levesque's seminal *BDI* logic. Using action types at the level of both syntax and semantics, Lorini and Herzig formalized the notions of successful and unsuccessful *attempts* to perform an action type. With the additional modalities for belief and for goals, the authors thereby integrated a detailed account of future- and present-directed intentions, where the execution of any such intention is an intentional-action execution.[11]

This concludes my discussion on the main topics in philosophy of intention that this chapter involves, as well as on previous logic-based formalizations of such topics. Now we are ready to proceed to my logic-based formalism, whose aforementioned goal is to represent the interplay between agency, knowledge, intentions, and intentional actions on the road to building a theory of responsibility.

5.3 My Proposal for a Logic of Intentionality

To address the challenge of incorporating a notion of intentionality—in terms both of mental states (intentions) and of modes of acting (intentional actions)—into the stit-theoretic conception of agency, I use *present-directed intentions*, written 'p-d

[9] Neither of the two works mentioned in this paragraph offered a sound and complete proof system for their corresponding logics.

[10] Broersen did present a sound and complete proof system for his logic of intentional action.

[11] Lorini and Herzig also presented a sound and complete proof system for their rich logic.

intentions' from here on. In stit theory, p-d intentions are best understood as those intentions that an agent has exactly at the moment of acting, right before making its choice, that concern states of affairs that are possible at precisely that moment. The instantaneous nature of agency in atemporal *BST* is what promotes the use of these p-d intentions to reason about the goals and plans that an agent has at the moment at which the agent is performing an action, where such goals and plans are about the states of affairs of that very moment. Still, perhaps a better way to think about p-d intentions is as the condensation of previous future-directed intentions in making a particular choice.

Therefore, here I extend *EST* with a modality of the form $I_\alpha\varphi$, meant to express that at an index agent α had a p-d intention toward the realization of φ. As for the semantics of $I_\alpha\varphi$, the idea is to assign a special topology to each agent. The non-empty open sets in any such associated topology will represent the agent's p-d intentions at the moment of acting, so that if a non-empty open set U in the topology associated with α supports φ (i.e., if φ holds at all indices within U), then U is a p-d intention of α toward the realization of φ. Roughly speaking, then, my proposal for the semantics of $I_\alpha\varphi$ is as follows: $I_\alpha\varphi$ holds at an index iff at such an index there exists U in the topology associated with α such that $U \subseteq \varphi$. In turn, the conjunction $[\alpha]\varphi \wedge I_\alpha[\alpha]\varphi$ is meant to evoke that α has intentionally seen to it that φ.[12] As mentioned in the introduction, I refer to the resulting logic as *intentional epistemic stit theory* (*IEST*).

Instead of spinning around the concepts informally, let me dive into the rigorous definitions right away.

5.3.1 Topologies of Intentions

I start by addressing some basic definitions from General Topology. For any other basic definitions that I might be taking for granted, the reader is referred to Willard (2004) or Engelking (1989) as proper background textbooks.

Definition 5.1 (Topological spaces)**.** *Let X be a set. Then $\tau \subseteq 2^X$ is called a* topology *on X if it meets the following requirements:*

- $X, \emptyset \in \tau$.
- Closure under finite intersections*: If $U, V \in \tau$, then $U \cap V \in \tau$.*
- Closure under arbitrary unions*: For a family $\mathcal{G} \subseteq \tau$, $\bigcup \mathcal{G} \in \tau$.*

[12] Of course, this reading of $I_\alpha\varphi$ and of the conjunction $[\alpha]\varphi \wedge I_\alpha[\alpha]\varphi$ positions my proposal as belonging to a particular philosophical standpoint on the relation between intentions and intentional action. I address the details of such a standpoint in Section 5.4.

A topological space*, then, is a pair* (X, τ) *such that* X *is a set and* τ *is a topology on* X*. The elements of* τ *are called* open sets*. Complements of open sets are called* closed sets*. For* $A \subseteq X$*, the* interior of A*, denoted by* $int(A)$*, is defined as the* $\subseteq$*-largest open set included in* A*. In turn, the* closure of A*, denoted by* $Cl(A)$*, is defined as the* $\subseteq$*-least closed set including* A*.*

Let $x \in X$ and $A \subseteq X$. Two standard results in General Topology are the following: (1) $x \in int(A)$ iff there exists an open set U such that $x \in U \subseteq A$; and (2) $x \in Cl(A)$ iff every open set U such that $x \in U$ intersects A ($U \cap A \neq \emptyset$).

Definition 5.2 (Density)**.** *For a topological space* (X, τ) *and* $A \subseteq X$*,* A *is said to be* τ-dense in X *iff* $Cl(A) = X$*, or, equivalently, iff for every non-empty open set* $O \in \tau$*,* $O \cap A \neq \emptyset$*.*

With these basic definitions, let me introduce the semantics for formulas of a language that extends *EST* (see Chapter 2's Subsection 2.4.4) with modality $I_\alpha\varphi$.

Definition 5.3 (Syntax of *IEST*)**.** *Given a finite set Ags of agent names and a countable set of propositions P, the grammar for the formal language* $\mathcal{L}_I$ *is given by*

$$\varphi ::= p \mid \neg\varphi \mid \varphi \wedge \varphi \mid \Box\varphi \mid [\alpha]\varphi \mid K_\alpha\varphi \mid I_\alpha\varphi,$$

where p ranges over P and α *ranges over Ags.*

In this language, $\Box\varphi$, $[\alpha]\varphi$, and K_α have the same meanings as in *EST* (Definition 2.26, p. 70); $I_\alpha\varphi$, in turn, expresses that 'agent α had a p-d intention toward the realization of φ,' or that 'α p-d intended φ,' or that 'α p-d intended that φ would hold.'[13] As for the semantics, the structures on which the formulas of $\mathcal{L}_I$ are evaluated are based on what I call *intentional epistemic branching-time frames*.

Definition 5.4 (*Iebt*-frames & models)**.** *A tuple* $\left\langle M, \sqsubset, Ags, \textbf{Choice}, \{\sim_\alpha\}_{\alpha \in Ags}, \tau \right\rangle$ *is called an* intentional epistemic branching-time frame *(*iebt*-frame for short) iff*

- $\left\langle M, \sqsubset, Ags, \textbf{Choice}, \{\sim_\alpha\}_{\alpha \in Ags} \right\rangle$ *is an* ebt*-frame (Definition 2.27, p. 70) that additionally satisfies the following conditions:*

[13] Just as in all the other chapters of this thesis, the present description of the stit-theoretic modalities follows *my interpretation* of the semantics (see the discussion on p. 34 and Remark 2.4, p. 36). Therefore, when specifying the points of evaluation for the formulas—the indices in *bt*-models—I take it that at those indices states of affairs are definitive. Because of this, I use the present-perfect tense for the description of modality $[\alpha]\varphi$ and say that 'at index $\langle m, h\rangle$ α has seen to it that φ.' To be consistent, I use the past tense for modalities $\Box\varphi$, $K_\alpha\varphi$, and $I_\alpha\varphi$ and say that 'at index $\langle m, h\rangle$ φ was settled,' that 'at index $\langle m, h\rangle$ α knew φ,' and that 'at index $\langle m, h\rangle$ α p-d intended φ.' As discussed in Chapter 2, this usage does not mean to refer to past moments. Rather, it aims to reinforce the notion that, at the level of indices, circumstances in the world are definitive, have already happened, and cannot be changed.

- $(\texttt{OAC})$ Own action condition*: for all $\alpha \in Ags$ and each index $\langle m,h\rangle$, $\langle m,h\rangle \sim_\alpha \langle m,h'\rangle$ for every $h' \in \mathbf{Choice}^m_\alpha(h)$.*
- $(\texttt{Unif}-\texttt{H})$ Uniformity of historical possibility*: for all $\alpha \in Ags$ and each index $\langle m,h\rangle$, if $\langle m,h\rangle \sim_\alpha \langle m',h'\rangle$, then for all $h_* \in H_m$ there exists $h'_* \in H_{m'}$ such that $\langle m,h_*\rangle \sim_\alpha \langle m',h'_*\rangle$.*

For $\alpha \in Ags$, two notions of α's information set at $\langle m,h\rangle$ are defined: the set $\pi^\Box_\alpha[\langle m,h\rangle] := \{\langle m',h'\rangle ; \exists h'' \in H_{m'}$ s. t. $\langle m,h\rangle \sim_\alpha \langle m',h''\rangle\}$ is α's ex ante information set at $\langle m,h\rangle$*; and the set $\pi_\alpha[\langle m,h\rangle] := \{\langle m',h'\rangle ; \langle m,h\rangle \sim_\alpha \langle m',h'\rangle\}$ is α's ex interim* information set at $\langle m,h\rangle$.[14]

- *τ is a function that assigns to each $\alpha \in Ags$ and index $\langle m,h\rangle$ a topology $\tau^{\langle m,h\rangle}_\alpha$ on $\pi^\Box_\alpha[\langle m,h\rangle]$. This is the* topology of α's intentionality at $\langle m,h\rangle$*, where any non-empty open set is interpreted as a p-d intention of α at $\langle m,h\rangle$. Additionally, τ must satisfy the following conditions:*

  - $(\texttt{CI})$ Finitary consistency of intention*: for all $\alpha \in Ags$ and each index $\langle m,h\rangle$, every non-empty $U, V \in \tau^{\langle m,h\rangle}_\alpha$ are such that $U \cap V \neq \emptyset$. In other words, every non-empty $U \in \tau^{\langle m,h\rangle}_\alpha$ is $\tau^{\langle m,h\rangle}_\alpha$-dense.*
  - $(\texttt{KI})$ Knowledge of intention*: for all $\alpha \in Ags$ and each index $\langle m,h\rangle$, if $\pi^\Box_\alpha[\langle m,h\rangle] = \pi^\Box_\alpha[\langle m',h'\rangle]$, then $\tau^{\langle m,h\rangle}_\alpha = \tau^{\langle m',h'\rangle}_\alpha$. In other words, α has the same topology of p-d intentions at all indices lying within α's current* ex ante *information set.*

An iebt*-model $\mathcal{M}$, then, results from adding a valuation function $\mathcal{V}$ to an* iebt*-frame, where $\mathcal{V} : P \to 2^{I(M\times H)}$ assigns to each atomic proposition a set of indices.*

Iebt-models allow us to provide semantics for the formulas of $\mathcal{L}_\mathsf{I}$:

Definition 5.5 (Evaluation rules for *IEST*). *Let $\mathcal{M}$ be an* iebt*-model. The semantics on $\mathcal{M}$ for the formulas of $\mathcal{L}_I$ are recursively defined as in Definition 2.28 (p. 70), with the following additional clause:*

$$\mathcal{M}, \langle m,h\rangle \models I_\alpha\varphi \quad \textit{iff} \quad \text{there exists } U \in \tau^{\langle m,h\rangle}_\alpha \text{ s. t. } U \neq \emptyset \text{ and } U \subseteq ||\varphi||,$$

where $||\varphi||$ denotes the set $\{\langle m,h\rangle \in I(M\times H); \mathcal{M}, \langle m,h\rangle \models \varphi\}$.

[14] This chapter adopts the same conventions, with respect to the topics that are relevant in *EST* (*knowingly doing, epistemic sense of ability, knowledge across the stages of information disclosure,* and *uniformity*), as Chapter 4 (see Subsection 4.2.2).

Therefore, one says that at index $\langle m,h\rangle$ agent α p-d intended φ iff there exists $U \in \tau_\alpha^{\langle m,h\rangle}$ that supports φ. As for intentional action, I characterize this notion with the conjunction $[\alpha]\varphi \wedge I_\alpha[\alpha]\varphi$, so that at index $\langle m,h\rangle$ agent α has intentionally seen to it that φ iff α has seen to it that φ and α p-d intended to see to it that φ.

5.3.1.1 Discussion

It is important to emphasize that, for each $\alpha \in Ags$ and index $\langle m,h\rangle$, the topology $\tau_\alpha^{\langle m,h\rangle}$ is a topology on α's *ex ante* information set. Thus, the logic *IEST* satisfies what I call the *knowledge-to-intention property*: at an index an agent's p-d intentions are information that was available to the agent regardless of anyone's choice of action (including the agent's one). Frame condition *knowledge of intention* (KI), then, implies a second important feature, which I refer to as the *knowledge-of-intention property*: at an index an agent always knew *ex ante* its p-d intentions. The arguments in favor of these two properties are given below.

- The *knowledge-to-intention property* concerns the fact that all p-d intentions are included in an agent's *ex ante* information set. To clarify, this property is reflected by the validity of formula $\Box K_\alpha\varphi \rightarrow I_\alpha\varphi$. As for arguments in favor of this property, Broersen (2011b) stated that intentions should be based on indices that an agent considers to be epistemically possible. Furthermore, it is reasonable to assume that if an agent knows φ *ex ante*, "which means the agent cannot do anything about it, the agent cannot but intend that φ holds" (Broersen, 2011b, p. 515).[15]

  Now, the idea that an agent cannot but intend everything known *ex ante* leads us to a notion of *non-deliberative intention*, so that an agent non-deliberatively p-d intends φ iff the agent knows φ *ex ante* (see the discussion on deliberative agency in Chapter 2's Section 2.2.5). Non-deliberative intentions thus concern states of affairs that the agent is compelled to intend for the sole reason that the agent knows that these states will ensue no matter what all agents do. In such terms, the p-d intentions underlying modality $I_\alpha\varphi$ can be thought of as *possibly non-deliberative p-d intentions*. In turn, a modality $I_\alpha^d\varphi$ for *deliberative intention*, intuitively expressing what agent α has chosen to intend, can be defined by setting $I_\alpha^d\varphi := I_\alpha\varphi \wedge \neg\Box K_\alpha\varphi$, so that at an index an agent deliberatively p-d intended φ iff at the index the agent p-d intended φ and the agent did not know φ *ex ante*. In turn, deliberative intentional

[15] Observe that α's *ex ante* knowledge, at a given index, is itself a p-d intention of α, as witnessed by the fact that, since $\tau_\alpha^{\langle m,h\rangle}$ is a topology on α's *ex ante* information set, then such an information set must be an element of the topology.

action can be characterized with the conjunction $[\alpha]\varphi \wedge I^d_\alpha[\alpha]\varphi$, so that at an index an agent has deliberative-intentionally seen to it that φ iff at the index the agent has intentionally seen to it that φ and it was not settled that the agent knowingly saw to it that φ.

- The *knowledge-of-intention property* concerns the fact that at an index an agent always knew *ex ante* its p-d intentions. To clarify, this property is reflected by the validity of formulas $I_\alpha\varphi \rightarrow \Box K_\alpha I_\alpha\varphi$ and $\neg I_\alpha\varphi \rightarrow \Box K_\alpha\neg I_\alpha\varphi$. This is a desirable property in virtue of a usual assumption of positive and negative introspection about one's own intentionality. According to Lorini and Herzig (2008), who formalized the relation between intentions and beliefs, agents have positive and negative introspection about their intentions with respect to their belief (see also Dunin-Keplicz & Verbrugge, 2002; Herzig & Longin, 2004). This means that formulas corresponding to $I_\alpha\varphi \rightarrow B_\alpha I_\alpha\varphi$ and to $\neg I_\alpha\varphi \rightarrow B_\alpha\neg I_\alpha\varphi$ are valid in their logics. Broersen (2011b) supported this claim and took it further so as to include positive and negative introspection about one's own intentions with respect to knowledge. Presently, I enforce this property—with respect to *ex ante* knowledge—following a relatively undisputed premise in philosophy of intention that Setiya (2011) referred to as *practical self-knowledge*. According to this premise, if an agent has the capacity to act for reasons and can ascribe intentions to others, then the agent has the capacity for (groundless) knowledge of its own intentions.[16]

Now, the fact that p-d intentions are dense implies that they are consistent, so that an agent cannot p-d intend both φ and $\neg\varphi$ at the same time. Commonly accepted in the philosophical literature on intentions (see, for instance, Bratman, 1987; Broersen, 2011b; Cohen & Levesque, 1990; Herzig & Longin, 2004; Lorini & Herzig, 2008), this property is reflected by the validity of schema (D) for $I_\alpha\varphi$. For agent α and state of affairs φ, $int\,(||\varphi||)$ represents the $\subseteq$-biggest p-d intention that α currently has toward the realization of φ. Observe, then, that either $int\,(||\varphi||) = \emptyset$ or $int\,(||\varphi||)$ intersects all p-d intentions of α at an index, and that $\mathcal{M}, \langle m, h\rangle \models I_\alpha\varphi$ iff $Cl\,(int\,(||\varphi||)) = \pi^\Box_\alpha[\langle m, h\rangle]$.

Before presenting some examples, I want to justify my choice of using topologies, since the reader might be curious about the reason for such a choice. As mentioned before, the intuition that an intention can be seen as a set of possibilities is fairly standard in the logic-based literature on intentions. Inspired by

[16] Setiya (2011, pp. 189–190) observed that "just as it is impossible for a subject with the power of inference and the concept of belief to lack first-person access to his own beliefs, so it is impossible for an agent who does things for reasons and has the concept of intention to lack first-person access to what she herself intends."

Konolige and Pollack's (1993) ideas behind using neighborhood semantics for formalizing intentionality (see also Duijf, 2018), I opted to represent p-d intentions as special subsets of indices in *bt*-models.[17] However, unlike these two approaches, I do not agree with the claim that all kinds of p-d intentions should not be closed under logical consequence. Thus, I started considering the idea of 'topologies of intentions' according to the intuition that unions and intersections of p-d intentions are also p-d intentions. After exploring different options for semantics, I opted for the topological representation because, in my view, it is straightforward, intuitive, and expressive, and because it helps in the description of a particular relation between agency, intentions, and intentional action. In *IEST*, an agent's p-d intentions at the moment of acting are the basis of intentional action insofar as there must be a p-d intention included in a choice that brings about φ in order for an agent to intentionally see to it that φ (recall that I presently characterize intentional action with the conjunction $[\alpha]\varphi \wedge I_\alpha[\alpha]\varphi$).[18]

5.3.1.2 Examples

To illustrate my semantics of intentionality, I present a formal analysis of two examples using *IEST*.

Example 5.6. *Recall the situation described at the beginning of this chapter, where you are a lawyer in the prosecution of a driver that ran over—and killed—a traffic officer. Consider the* iebt*-model* $\mathcal{M}$ *depicted in Figure 5.1.*

Here, $Ags = \{\text{driver}\}$*, and* m_1 *is a moment. There are two histories (*h_1 *and* h_2*) passing through* m_1*. At* m_1 *the choices of action available to* driver *are the following:* R_1*, standing for the choice of running over the traffic officer, and* R_2*, standing for the choice of stopping the car. According to the choice performed, time moves on either into history* h_1 *or into history* h_2*. As implied by the statement of the example,* h_1 *is the actual history.*

[17] As pointed out by Pacuit (2007) in his lecture notes for a course on neighborhood semantics, "[s]ets paired with a distinguished collections of subsets are ubiquitous in many areas of mathematics."

[18] The reader might wonder why I did not opt for a relational semantics for intentionality. As mentioned here, I think that the topological representation is better suited to the notion of p-d intentions than potential relational representations. To be sure, the relational paradigms that I explored before choosing topology involved a more convoluted truth condition for $I_\alpha\varphi$, as well as a less clear-cut formulation of my models' constraints. Moreover, topological semantics generalize standard relational semantics and are more expressive. In the words of Özgün (2017, Chapter 1, p. 2), "topological spaces are equipped with well-studied basic operators such as the interior and closure operators which, alone or in combination with each other, succinctly interpret different modalities, giving a better understanding of their axiomatic properties." As implied by the discussion on the problems in philosophy of intention in Section 5.2, it is not easy to model the relation between the concepts of future-directed intentions, intentional action, and intention-with-which. In instantaneous stit theory—without temporal modalities—such a relation is even harder to address. The topological semantics for p-d intentions, then, helped me establish one such relation in clear, tractable terms.

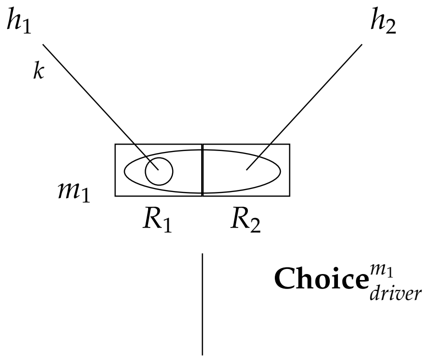


Figure 5.1: *Driver example.*

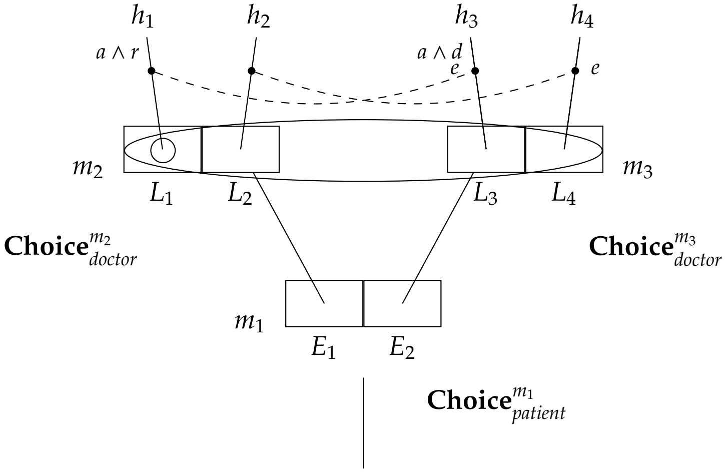
Throughout the diagrams of this thesis, I have represented the epistemic states of agents by indistinguishability relations given by dashed lines, omitting reflexive loops. Here, there are no dashed lines because driver *is assumed to distinguish* $\langle m_1, h_1\rangle$ *from* $\langle m_1, h_2\rangle$. *Thus, at every index based on* m_1 driver *knew her choice of action. In turn,* driver*'s* ex ante *information set at the actual index* $\langle m_1, h_1\rangle$*—denoted by* $\pi^{\Box}_{\text{driver}}[\langle m_1, h_1\rangle]$*—is the set* $\{\langle m_1, h_1\rangle, \langle m_1, h_2\rangle\}$*, which coincides with* $\pi^{\Box}_{\text{driver}}[\langle m_1, h_2\rangle]$.

As for driver*'s intentionality, consider the topology* $\tau^{\langle m_1,h_1\rangle}_{\text{driver}}$. *Since* $\pi^{\Box}_{\text{driver}}[\langle m_1, h_1\rangle] = \pi^{\Box}_{\text{driver}}[\langle m_1, h_2\rangle]$*, frame condition* (KI) *implies that* $\tau^{\langle m_1,h_1\rangle}_{\text{driver}} = \tau^{\langle m_1,h_2\rangle}_{\text{driver}}$. *The non-empty open sets of such a topology are represented in the diagram using circles and ellipses. More precisely,* $\tau^{\langle m_1,h_1\rangle}_{\text{driver}} = \left\{\emptyset, \pi^{\Box}_{driver}[\langle m_1, h_1\rangle], \{\langle m_1, h_1\rangle\}\right\}$.

Let k stand for the atomic proposition 'the traffic officer has been killed' in Figure 5.1. Thus, according to the diagram, $\mathcal{M}, \langle m_1, h_1\rangle \models K_{driver}[driver]k$, for instance: *at the actual index,* driver *knowingly killed the traffic officer*.

To illustrate the evaluation of formulas involving *driver*'s intentionality, let U denote the set $\{\langle m_1, h_1\rangle\}$. Then $U \in \tau^{\langle m_1,h_1\rangle}_{driver}$ and $U \subseteq ||k||$. Thus, according to Definition 5.5, $\mathcal{M}, \langle m_1, h_1\rangle \models I_{driver}k$: *at the actual index* driver *had a p-d intention—or p-d intended—that the officer was killed*. The same U attests to the fact that $\mathcal{M}, \langle m_1, h_1\rangle \models I_{driver}[driver]k$: *at the actual index* driver *had a p-d intention to see to it that the officer was killed*. As such, for all practical purposes, *driver* knowingly and intentionally killed the officer—which makes it reasonable for her to be blamed for second- or first-degree murder.

Example 5.7. *Recall Chapter 3's Example 3.12 (p. 112). This example involves a doctor who supplied anesthetics to a patient before a surgery. The patient had eaten just before* *the surgery, and the doctor did not know this. Anesthetics must have been supplied only on an empty stomach, so the patient died due to the interaction between the food and the anesthetics. Consider the* iebt*-model* $\mathcal{M}$ *depicted in Figure 5.2.*



h_1
h_2
h_3
h_4
$a \wedge r$
$a \wedge d$
e
e
m_2
m_3
L_1
L_2
L_3
L_4
$\mathbf{Choice}^{m_2}_{doctor}$
$\mathbf{Choice}^{m_3}_{doctor}$
m_1
E_1
E_2
$\mathbf{Choice}^{m_1}_{patient}$

Figure 5.2: *Anesthesiologist example, again.*

Here, $Ags = \{\text{patient}, \text{doctor}\}$*, and* m_1*,* m_2*, and* m_3 *are moments, where* $\sqsubset$ *is defined so as to be represented by the diagram. There are four histories* (h_1–h_4)*, representing different possibilities for time to evolve according to the actions available both to* patient *and* doctor*. At* m_1 *we find two choices available to* patient*:* E_1*, standing for the choice of refusing to eat, and* E_2*, standing for the choice of eating. According to the action chosen by* patient*, the world evolves toward either* m_2 *or* m_3*. At both these moments, it is* doctor*'s turn to act, and her available choices are the following:* L_1 *and* L_3*, standing for supplying anesthetics; and* L_2 *and* L_4*, standing for refusing to supply anesthetics. As implied by the statement of the example,* h_3 *is the actual history.*

The epistemic states that I focus on are those of doctor*. They are represented with the indistinguishability relation given by dashed lines (where reflexive loops are omitted). Thus, at all indices based on* m_2 *and* m_3 doctor *did not know whether the patient had eaten. However, at such indices she did know which action she performed.*

As for doctor's intentionality, let me present $\tau^{\langle m_3,h_3\rangle}_{doctor}$*. Observe that, for all* $i \in \{2,3\}$ *and* $h \in H_{m_i}$*,* $\pi^{\Box}_{doctor}[\langle m_i,h\rangle] = \left\{\left\langle m_j,h'\right\rangle; j \in \{2,3\} \text{ and } h' \in H_{m_j}\right\}$*. This implies that* $\tau^{\langle m_2,h_k\rangle}_{doctor} = \tau^{\langle m_3,h_l\rangle}_{doctor}$ *for all* $k \in \{1,2\}$ *and* $l \in \{3,4\}$*. Once again, the non-empty open sets of such a topology are represented in the diagram using circles and ellipses. Therefore,* $\tau^{\langle m_3,h_3\rangle}_{doctor} = \left\{\emptyset, \pi^{\Box}_{doctor}[\langle m_3,h_3\rangle], \{\langle m_2,h_1\rangle\}\right\}$.

Let e stand for the atomic proposition 'the patient has eaten' in Figure 5.2, and let a stand for 'anesthetics are supplied to the patient,' r stand for 'the patient is ready for surgery,' and d stand for 'the patient will die.' Then consider the following examples for the evaluation of formulas of $\mathcal{L}_I$: $\mathcal{M}, \langle m_3, h_3\rangle \models [doctor]d$: *at the actual index* doctor *has seen to it that the patient will die*; and $\mathcal{M}, \langle m_3, h_3\rangle \models \neg K_{doctor}[doctor]d$: *at the actual index* doctor *did not knowingly kill the patient*.

As for formulas involving *doctor*'s intentionality, let $U = \{\langle m_2, h_1\rangle\}$. Then $U \in \tau_{doctor}^{\langle m_3,h_3\rangle}$ and $U \subseteq ||a \wedge r||$. Thus, $\mathcal{M}, \langle m_3, h_3\rangle \models I_{doctor}(a \wedge r)$: *at the actual index* doctor *had a p-d intention that the anesthetics would be supplied and that the patient would get ready for surgery*. Similarly, observe that $\mathcal{M}, \langle m_3, h_3\rangle \models \neg I_{doctor}d$: *at the actual index* doctor *had no p-d intention that the patient would die*, which implies that $\mathcal{M}, \langle m_3, h_3\rangle \models \neg I_{doctor}[doctor]d$: *at the actual index* doctor *did not have a p-d intention to kill the patient and thus did not intentionally kill him*.

Therefore, the model tells us two important facts: (1) although *doctor* killed the patient on the causal level, she acted neither knowingly nor intentionally; and (2) *doctor* actually p-d intended that the patient would live. These claims provide good reasons for excusing the doctor from having moral responsibility of the patient's death.

5.4 Logic-Based Properties & Axiomatization

Let me present and discuss some properties of *IEST*, in terms of formulas that are either valid or invalid with respect to *iebt*-models.

5.4.1 Properties

The logic-based properties of modalities $\Box\varphi$ and $[\alpha]\varphi$ are the same as those reviewed in Chapter 2's Subsection 2.3.1. The properties of knowledge and its interplay with agency are the same as those addressed in Chapter 4's Subsection 4.5.1: K_α is an **S5** operator such that the formulas associated with frame conditions $(\mathtt{OAC})$ and $(\mathtt{Unif-H})$ are valid (see items 5 and 6 in the list of *EAUST*'s logic-based properties, Chapter 4, Subsection 4.5.1, pp. 172 and 173).

As for operator I_α, it is a **KD** operator. The validity of the **KD** schemata for I_α follows from Definitions 5.4 and 5.5, and it has the following consequences for my notion of intentionality:

- The validity of (K) $(I_\alpha(\varphi \to \psi) \to (I_\alpha\varphi \to I_\alpha\psi))$ implies that if at an index an agent p-d intended φ then the agent p-d intended all the logical consequences of φ. Thus, my notion of (possibly non-deliberative) p-d intentions is vulnerable to a particular version of the so-called *side-effect problem* (see Footnote 8). I do not agree with the claim that possibly non-deliberative intentions should not be closed under logical consequence. The reason is that stit-theoretic agents are rational, logically omniscient thinkers, who know *ex ante* all the logical consequences of what they know. Similarly, they know *ex ante* all tautologies and all valid formulas. For agents of this kind, I find it reasonable to assume that they will p-d intend the logical consequences of whatever they intend. Furthermore, formula $(\Box K_\alpha(\varphi \to \psi) \wedge I_\alpha\varphi) \to I_\alpha\psi$ is valid in my framework, so that if at an index an agent knew *ex ante* that φ implies ψ—or non-deliberatively p-d intended that φ implies ψ—then the agent's p-d intention of φ implies its p-d intention of ψ. Thus, p-d intentions are presently closed under *ex ante* knowledge (or non-deliberative p-d intention) of side-effect implication. Again, I find that this is a reasonable assumption for rational, logically omniscient thinkers.[19]

Philosophers might remain skeptical about this line of argumentation. Observe, then, that my framework admits a solution to the versions of the side-effect problem discussed above. It is easy to verify that the concept of deliberative intention (p. 228), expressed by modality $I^d_\alpha\varphi$, is not closed under logical consequence nor under *ex ante* knowledge of side-effect implication. Reminiscent of what happens in Cohen and Levesque's (1990) and Broersen's (2011b) proposals for formalizing intentions that are not closed under logical consequence, however, if the agent did not already know *ex ante* that the

[19] Most versions of the side-effect problem (see, for instance, Bratman, 1987; Broersen, 2011b; Cohen & Levesque, 1990; Rao & Georgeff, 1991) only argue that intentions should not be closed under *believed* (or anticipated) side effects, in the sense that if an agent intends φ and also believes that $\varphi \to \psi$, then one should not conclude that the agent also intends ψ. In fact, the formulation of the side-effect problem with the dentist example (Footnote 8) involves only this argument, in terms of intentions and belief. Indeed, it is unclear to me why the side-effect problem is sometimes assumed to refer to closure of intentions under logical consequence (Duijf, 2018; Konolige & Pollack, 1993). Logical consequence is a very strong assumption for most logic-based models, since it means that $\varphi \to \psi$ holds *at all states*. For logically ideal agents as the ones here modelled, this implies that at every possible configuration of the world every agent knows for sure (with absolute, indefeasible certainty) that $\varphi \to \psi$ will hold (at every possible configuration of the world). Recall once again the dentist example (Footnote 8), and suppose that we phrase it in terms of logical consequence. Now pain is a necessary consequence of getting one's tooth filled at every possible configuration of the world, and the agent not only believes that the process of having his tooth filled will cause him pain, but he also knows for sure that there is no possible configuration of the world in which he will not feel pain by getting his tooth filled. Can we really say that the agent intended to get his tooth filled without intending pain? Most likely, this is why Rao and Georgeff (1991), Lorini and Herzig (2008), and Bentzen (2012), for instance, all disregard that intentions should be not closed under logical consequence. As for approaches that manage to yield logics of intending where intentions are not closed under logical consequence, it is either the case that non-closure is only possible when side effects are already known/believed (Broersen, 2011b; Cohen & Levesque, 1990) or the case that the logics have problems arising from the use of non-normal operators (Duijf, 2018; Konolige & Pollack, 1993).

side effect held then the agent also has a deliberative p-d intention of the side effect. To clarify, formula $\left(I^d_\alpha\varphi \wedge \Box K_\alpha(\varphi \to \psi) \wedge \neg\Box K_\alpha\psi\right) \to I^d_\alpha\psi$ is still valid.

- The validity of (D) ($I_\alpha\varphi \to \neg I_\alpha\neg\varphi$) implies that if at an index an agent p-d intended φ then at that index the agent must not have p-d intended $\neg\varphi$. As mentioned before, most of the authors whose formalization of intention has been discussed (Bratman, 1987; Broersen, 2011b; Cohen & Levesque, 1990; Herzig & Longin, 2004; Lorini & Herzig, 2008) support the idea that, at a specific point in time, future-directed intentions, p-d intentions, intentional actions, and intentions-with-which should be respectively consistent, and I agree with them.

Furthermore, the validity, resp. invalidity, of the following formulas, with respect to the class of *iebt*-models, captures important properties of the interplay between the modalities of *IEST*.

1. (a) $\not\models I_\alpha\varphi \to I_\alpha[\alpha]\varphi$: it is not necessarily true that if at an index an agent p-d intended φ then at that index the agent p-d intended to see to it that φ. This property refers to a distinction between intending that φ is the case and intending to be the material author of φ. For instance, suppose that I am a dictator displaying psychopathic traits. I have an intention toward the bombing of a neighboring country, but I do not intend for me to actually press the button that would deploy a bomb. Although some authors claim that the most primal notion of intending always refers to *intending to do* (see, for instance, Moran & Stone, 2009; Thompson, 2008), I support the idea—consistent with Bratman's (1984) seminal thesis that future-directed intentions are elements in complex plans—that an agent can intend the realization of some state of affairs without intending to be the one realizing it.[20] Once again, a good example of

[20] Duijf (2018, Chapter 4, p. 163) explicitly stated that there is a distinction between intending and intending to do. He wrote: "[t]here are two different types of future-directed intentions: I can intend to perform a certain action, or I can intend to realize a certain state of affairs." Observe that I have often used expressions such as 'α p-d intended that φ would hold,' or 'α p-d intended φ' to describe modality $I_\alpha\varphi$, reserving expressions of the form 'α p-d intended to do φ,' or 'α p-d intended to see to it that φ,' to describe the combined modality $I_\alpha[\alpha]\varphi$. This presupposes a practical identification of *α's intending to do something* with *α's intending that something will be done by α.* Therefore, my framework has two points of contention with Thompson's (2008)'s view (which in turn follows the tradition that began with Anscombe's (1963) work):

i. The logic-based property that the present footnote annotates ($\not\models I_\alpha\varphi \to I_\alpha[\alpha]\varphi$) implies that I distinguish between intentions, on the one hand, and intending to do, on the other. In contrast, Thompson formalized a sense of intention in which intending is already intending to do, and furthermore already an intentional action "in progress" (Setiya, 2018).

this lies in mastermind agents that delegate actions to subordinates. The distinction between intending that φ is the case, on the one hand, and intending to actually see to it that φ, on the other, is all the more relevant in responsibility attribution: although my subordinate pilots were the ones deploying the bombs, it is me who should stand trial in The Hague.

To illustrate this property, consider a variation of Example 5.6. Suppose that *driver* did not want to run over the traffic officer herself, but, still, she had a p-d intention that the officer would get killed. A diagram of this situation is included in Figure 5.3. In this case, observe that

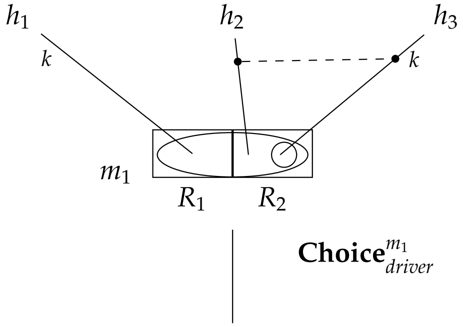


Figure 5.3: *Another driver example.*

$\tau^{\langle m_1,h_3\rangle}_{driver} = \left\{\emptyset, \pi^{\Box}_{driver}[\langle m_1,h_3\rangle], \{\langle m_1,h_3\rangle\}\right\}$. Let $U = \{\langle m_1,h_3\rangle\}$. Then $U \subseteq ||k||$. This means that $\mathcal{M}, \langle m_1,h_3\rangle \models I_{driver}k$: *at* $\langle m_1,h_3\rangle$ driver *p-d intended that the officer would get killed.* However, there does not exist a non-empty open set included in $||[driver]k||$, which means that $\mathcal{M}, \langle m_1,h_3\rangle \models \neg I_{driver}[driver]k$: *at* $\langle m_1,h_3\rangle$ driver *did not p-d intend to kill the officer.*

(b) $\not\models I_\alpha[\alpha]\varphi \to [\alpha]\varphi \wedge I_\alpha[\alpha]\varphi$: it is not necessarily true that if at an index an agent p-d intended to see to it that φ then at that index the agent has intentionally seen to it that φ. Recall that I interpret $[\alpha]\varphi \wedge I_\alpha[\alpha]\varphi$ as expressing that α has intentionally seen to it that φ. Thus, this property

[ii] The expressions here used to describe modalities $I_\alpha\varphi$ and $I_\alpha[\alpha]\varphi$ imply that, in my framework, it is possible to reduce *intending to do* to *intending that*. If $I_\alpha[\alpha]\varphi$ holds, my interpretation says that '*α intended that α* has seen to it that φ,' and I have identified such an expression with '*α intended to do φ*.' In contrast, Thompson (2008, Chapter 8, pp. 120–123) considered that there is a primal sense of intending that only takes verb phrases as complement. When somebody says 'I intend to walk to school,' for instance, this is not reducible to an expression of the form 'I intend that I will walk to school,' since the latter is not directed to a particular action of mine that will get me to school.

is in line with the assumption, that began with Davidson's (1978) work, that intending does not imply intentionally doing. For instance, I could have intended to start my car and still not have taken any action toward starting it.[21] To illustrate this property, consider Example 5.6. Here, $\mathcal{M}, \langle m_1, h_2 \rangle \models I_\alpha[driver]k$ and $\mathcal{M}, \langle m_1, h_2 \rangle \models \neg[driver]k$: *at* $\langle m_1, h_2 \rangle$ driver *p-d intended to kill the traffic officer, but at such an index* driver *did not intentionally kill the officer.*

(c) $\models I_\alpha[\alpha]\varphi \rightarrow I_\alpha\varphi$: if at an index an agent p-d intended to see to it that φ, then at that index the agent p-d intended φ. Since I interpret the conjunction $[\alpha]\varphi \wedge I_\alpha[\alpha]\varphi$ as α's intentionally doing φ, then this property yields that in *IEST* intentional action implies intending. Thus, my notion of intentionality falls under a philosophical standpoint that Bratman (1984) called the *Simple View*. The *Simple View* considers that, for an agent to intentionally do φ, the agent must also intend that φ is the case. Although Bratman heavily objected to the *Simple View*, I find it appropriate for logically omniscient agents.[22] The validity of this formula follows from the validity of schema (T) for $[\alpha]$, Necessitation for I_α, and the validity of schema (K) for I_α.

(d) $\not\models [\alpha]\varphi \rightarrow I_\alpha\varphi$: it is not necessarily true that if at an index an agent has seen to it that φ then at that index the agent p-d intended φ. This property reflects the desirable tenets that (i) not all actions follow a specific p-d intention, and that (ii) not all actions are intentional. As for

[21] The desirability of this property depends on agreeing with the tradition inspired by Davidson's work (see also Aune, 1977; Bratman, 1984, 1987; Searle, 1983). As mentioned in Section 5.2, the property would not be in line with what the camp that arose from the work of Anscombe (1963) thinks about the relation between intentions and intentional action.

[22] Bratman (1984) objected to the *Simple View* by showing that there are situations when one would naturally say that an agent intentionally did φ without actually intending φ. His famous example involves an agent playing two video games that are linked to each other. In each game, the objective is the same: guiding a missile to its respective target. The games are difficult, and the agent is doubtful of success at either of them. Moreover, the agent knows that the two games are linked in such a way that it is impossible to hit *both* targets. If both targets are about to be hit, simultaneously, then the machines just shut down. Both targets are visible to the agent, so the agent can see which target was hit, if any. Thus, the agent proceeds to try to hit target 1 and also to try to hit target 2, considering the risk of shutting down the machines as outweighed by the increase in the chances of hitting a target. Supposing that the agent hits target 1, it seems fair to say that the agent hit target 1 intentionally. So, on the *Simple View*, the agent must have intended to hit target 1. Symmetrically, the agent must also have intended to hit target 2. However, given the agent's knowledge that both targets cannot be hit, these two intentions are not consistent. Having them would involve the agent in a criticizable form of irrationality. However, it seems clear that the agent is not irrational in choosing the strategy of trying to hit both targets, because the games are both difficult. If the agent is not guilty of irrationality, then the agent should not be seen as having both intentions. Therefore, the *Simple View* should be false.

I find much worth in Bratman's (1984) solution to this problem of the *Simple View*. He introduced the notion of *motivational potential* of φ—referring to all ψ's which an agent counts as doing intentionally point (i), observe that agents can bring about states of affairs without any intention toward the realization of these states of affairs. This is what happens, for instance, in Example 5.7: *doctor* caused the patient's death, but *doctor* did not p-d intend that the patient would die. In terms of formulas, $\mathcal{M}, \langle m_3, h_3 \rangle \models [doctor]d \wedge \neg I_{doctor}d$. As for point (ii), it is clear that agents can bring about states of affairs unintentionally. Observe that the fact that $[\alpha]\varphi \rightarrow I_\alpha\varphi$ is not valid, coupled with the validity of the formula in item 1c ($I_\alpha[\alpha]\varphi \rightarrow I_\alpha\varphi$), implies that $[\alpha]\varphi \rightarrow I_\alpha[\alpha]\varphi$ is also not valid. Therefore, in light of my characterization of intentional action with the conjunction $[\alpha]\varphi \wedge I_\alpha[\alpha]\varphi$, the fact that $[\alpha]\varphi \rightarrow I_\alpha[\alpha]\varphi$ is not valid reflects that in *IEST* agents can act unintentionally.

2. (a) $\not\models K_\alpha\varphi \rightarrow I_\alpha\varphi$: it is not necessarily true that if at an index an agent knew φ then at that index the agent p-d intended φ. In light of the validity of the formulas associated with frame condition (OAC) (see item 5 in the list of *EAUST*'s logic-based properties, Chapter 4, Subsection 4.5.1, p. 172), $K_\alpha\varphi$ is logically equivalent to $K_\alpha[\alpha]\varphi$. To know φ, then, is to knowingly do φ (or to have *ex interim* knowledge of φ). Therefore, this property, which can be reformulated as $\not\models K_\alpha[\alpha]\varphi \rightarrow I_\alpha\varphi$, reflects the desirable tenet that knowingly doing φ does not imply intending φ, as can occur when someone else forced your hand, for instance. To clarify, consider another variation of Example 5.6. Suppose that *driver* did not want to run over the officer herself. By previously threatening to injure your family if you refused to follow her instructions, *driver* forced you into taking your own car and running over the officer. A diagram of your situation as an agent is included in Figure 5.4.

   In this case, the actual history is h_1. Thus, $\mathcal{M}, \langle m_1, h_1 \rangle \models K_{you}[you]k \wedge \neg I_{you}k$: *at the actual index* you *knowingly killed the traffic officer, but* you *had no p-d intention that the officer would be killed*. Furthermore, observe

in the course of carrying out an intention to do φ—and used the relation between this potential and beliefs/desires in complex plans to offer a solution to the *Simple View*'s problem. According to his proposal, in the video-games example, although the agent did not intend to hit target 1, hitting target 1 is included in the motivational potential of ψ = *getting a reward from the video-games*, for instance. Thus, although the agent does not intend to hit target 1, hitting it does count as an intentional action. Nevertheless, I disagree with the idea that the video-games example is suited to claiming that the *Simple View* is false. I do not think that, in such an example, the agent intentionally hit target 1. Let me explain this position, by means of my own framework. Let t_1 stand for the proposition 'target 1 is hit,' let t_2 stand for the proposition 'target 2 is hit,' and let α be the agent playing the video-games. In my view, formula $I_\alpha[\alpha]t_1$ does not hold, and thus α did not intentionally hit target 1. The formula that holds, rather, is $[\alpha](t_1 \wedge \neg t_2) \wedge I_\alpha[\alpha]((t_1 \wedge \neg t_2) \vee (\neg t_1 \wedge t_2))$. Thus, what the agent did intentionally was an exclusive disjunction: the agent intentionally either hit target 1 while not hitting target 2 or hit target 2 while not hitting target 1.

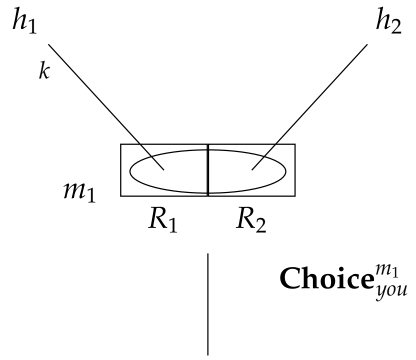


Figure 5.4: *Yet another driver example.*

that the validity of the formula in item 1c ($I_\alpha[\alpha]\varphi \to I_\alpha\varphi$) implies that $\mathcal{M}, \langle m_1, h_1 \rangle \models \neg I_{you}[you]k$ as well: *at the actual index* you *did not p-d intend to kill the traffic officer.* In light of my characterization of intentional action with the conjunction $[\alpha] \wedge I_\alpha[\alpha]\varphi$, this means that you did not kill the officer intentionally. Thus, in *IEST* knowingly doing also does not imply intentionally doing.

(b) $\not\models [\alpha]\varphi \wedge I_\alpha[\alpha]\varphi \to K_\alpha[\alpha]\varphi$: it is not necessarily true that if at an index an agent has seen to it that φ and the agent p-d intended to see to it that φ then at that index the agent has knowingly seen to it that φ. This property entails that my framework allows us to model situations where an agent intentionally does φ without knowingly doing φ. A good example of the viability of such situations is given by a small variation of one of Horty's (2019) three puzzles (the ones that were extensively discussed in Chapter 4's Sections 4.3 and 4.4). Suppose that *Nikolai* and *Dolokhov* are playing a game at a gambling house. The set-up is as follows: *Dolokhov* places a coin on top of a table—either heads up or tails up—and hides it from *Nikolai*. *Nikolai* can bet that the coin is heads up or bet that it is tails up. If *Nikolai* bets and chooses correctly, *Nikolai* and *Dolokhov* win 10 roubles from the house. If he chooses incorrectly, they win nothing. Assume that *Nikolai* p-d intended to win at any index based on m_2 and m_3. Then a diagram of the situation is depicted in Figure 5.5.

Just as in Chapter 4, at moment m_1 *Dolokhov* chooses between placing his coin on the table either heads up or tails up. Thus, his available actions are labelled by D_1 (placing the coin heads up) and D_2 (placing

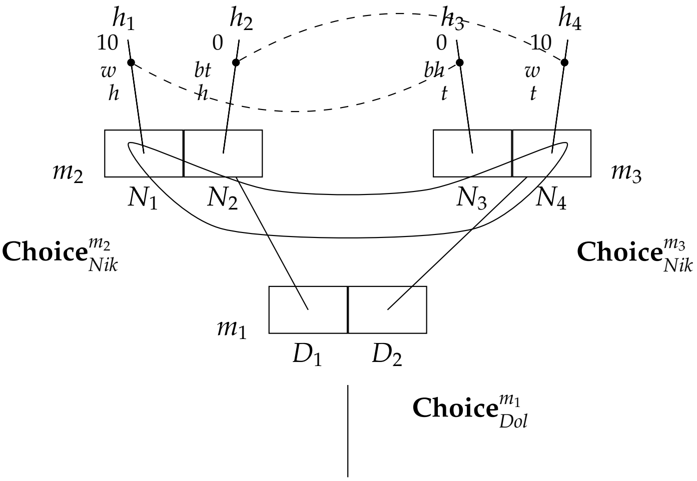


Figure 5.5: *Nikolai gambling.*

the coin tails up). At moments m_2 and m_3 it is *Nikolai*'s turn to act, and the available actions are the following: N_1 and N_3, where he bets heads; and N_2 and L_4, where he bets tails.

Here, h stands for the proposition 'Dolokhov's coin is placed heads up,' t stands for 'Dolokhov's coin is placed tails up,' bh stands for 'Nikolai has bet heads,' bt stands for 'Nikolai has bet tails,' and w stands for 'Nikolai and Dolokhov win.' Observe that *Nikolai*'s *ex ante* information set is the same at all indices based on m_2 and m_3: $\pi^{\Box}_{Nik}\left[\left\langle m_i, h_j\right\rangle\right] = \pi^{\Box}_{Nik}\left[\langle m_k, h_l\rangle\right]$ for all $i, k \in \{2, 3\}$ and j, l in 1–4. Let us assume that the actual index is $\langle m_2, h_1\rangle$, where *Nikolai* has bet heads and has won 10 roubles. The diagram shows that $\mathcal{M}, \langle m_2, h_1\rangle \models [Nik]w \wedge \neg K_{Nik}[Nik]w$: *at* $\langle m_2, h_1\rangle$*, although* Nikolai *has won the bet, he has done so unknowingly*. As for *Nikolai*'s intentionality, $\tau^{\langle m_2,h_1\rangle}_{Nik} = \left\{\emptyset, U, \pi^{\Box}_{Nik}\left[\langle m_2, h_1\rangle\right]\right\}$, where $U = \{\langle m_2, h_1\rangle, \langle m_3, h_4\rangle\}$.[23] Observe, then, that $U \subseteq \|[Nik]\, w\|$, so that $I_{Nik}[Nik]w$ holds at $\langle m_2, h_1\rangle$: *at* $\langle m_2, h_1\rangle$ Nikolai *p-d intended to win the bet*. Thus, *Nikolai* did not knowingly win the bet, but he did win it intentionally: $\neg K_{Nik}[Nik]w \wedge [Nik]w \wedge I_{Nik}[Nik]w$ holds.

The fact that my framework allows situations where an agent intentionally does φ without knowingly doing φ implies that it deviates from

[23] For the diagram's readability, I omitted displaying the ellipse that represents the full *ex ante* information set $\pi^{\Box}_{\alpha}\left[\langle m_2, h_1\rangle\right]$.

what Marcus (2019, p. 4) called the *knowledge thesis* for intentional action, according to which "it is impossible for a person to do something intentionally without knowing that she is doing it." Although the knowledge thesis is defended by many philosophers (see, for instance Anscombe, 1963; Broersen, 2011b; Gorr & Horgan, 1982; Olsen, 1969), there is empirical evidence that "in various scenarios a majority of non-specialists regard agents as intentionally doing things that the agents do not know they are doing and are not aware of doing" (Vekony, Mele, & Rose, 2021, p. 1231).[24]

(c) $\models \Box K_\alpha \varphi \rightarrow I_\alpha \varphi$: if at an index an agent knew φ *ex ante*, then at that index the agent (non-deliberatively) p-d intended φ. The validity of $\Box K_\alpha \varphi \rightarrow I_\alpha \varphi$ reflects what I called the *knowledge-to-intention property* in the discussion right after Definition 5.5, concerning the fact that all p-d intentions are included in an agent's *ex ante* information set. The arguments in favor of the *knowledge-to-intention property* were given on p. 228, and a proof of validity of $\Box K_\alpha \varphi \rightarrow I_\alpha \varphi$ follows from Definitions 5.4 and 5.5.

3. (a) $\models I_\alpha \varphi \rightarrow \Box K_\alpha I_\alpha \varphi$: if at an index an agent p-d intended φ, then at that index the agent knew *ex ante* that it p-d intended φ. Together with the validity of the formula in item 3b below, the validity of this one reflects what I called the *knowledge-of-intention property* in the discussion right after Definition 5.5, concerning the fact that at an index an agent must have known *ex ante* its p-d intentions. The arguments in favor of the *knowledge-of-intention property* were given on p. 229. As mentioned before, such a property is associated with frame condition (`KI`) in Definition 5.4. Indeed, formula $I_\alpha \varphi \rightarrow \Box K_\alpha I_\alpha \varphi$ defines (`KI`) insofar as an *ebt*-frame including τ satisfies (`KI`) iff $I_\alpha \varphi \rightarrow \Box K_\alpha I_\alpha \varphi$ is valid with respect to said frame. The validity of this formula follows from Definitions 5.4 and 5.5.

[24] Vekony et al. (2021) conducted two studies on groups of 250 people, asking them to rate the intentionality, knowledge, and awareness of an agent's actions in two different scenarios, where each study tested one of these scenarios. In the first study, the scenario involved a basketball player that is practicing free throws. One evening, he lines up and takes the shot, but just as the ball leaves his hands, lightning strikes the building. Due to this, the player is completely unaware of whether he sank the shot, although he did in fact sink the shot. In the second study, the scenario involved an agent that locks her door every morning as she leaves for work. On her way out to work one morning, she locks the door, but because she is preoccupied with thoughts about her day she is completely unaware of doing so. Therefore, she walks back from her car to check if she locked the door. For both studies, most participants considered that the agents were intentionally but unknowingly performing the actions of making the free throw and locking the door, respectively.

(b) $\models \neg I_\alpha \varphi \rightarrow \Box K_\alpha \neg I_\alpha \varphi$: if at an index an agent did not p-d intend φ, then at that index the agent knew *ex ante* that it did not p-d intend φ. See the discussion of item 3a above. In the proof system for *IEST* presented in Subsection 5.4.2, this formula can be derived using the one in item 3a above, so it is also valid (Observation 5.9 a).

Recall that in Section 5.2 I mentioned that two main problems in philosophy of intention are (1) the quest for unity in the three forms of intentionality (*future-directed intentions*, *intentional action*, and *intention-with-which*), and (2) the relation between intentionality, on the one hand, and knowledge, beliefs, desires, etc., on the other. The logic-based properties in item 1 somewhat settle where my interpretation of intentionality stands with respect to problem (1). In turn, the logic-based properties in items 2 and 3 speak of my take on a relation that is relevant in problem (2): the relation between intention and knowledge. Let me briefly elaborate on these matters.

As for problem (1), I prioritize p-d intentions—which lie in the same category as future-directed intentions—and base on them the notion of intentional action. Since I identify α's intentionally doing φ with the conjunction $[\alpha]\varphi \wedge I_\alpha[\alpha]\varphi$, then at a given index α has intentionally seen to it that φ only if α p-d intended to see to it that φ—that is, only if $I_\alpha[\alpha]\varphi$ holds; therefore, the validity of formula $I_\alpha[\alpha]\varphi \rightarrow I_\alpha\varphi$ (item 1c) implies that for α to intentionally do φ α must have p-d intended that φ would be the case. As mentioned before, this means that my treatment of intentionality falls under what Bratman (1984) referred to as the *Simple View*.[25]

As for problem (2), my framework's position on the relation between an agent's intentionality and its knowledge can be summarized with two remarks. First, an agent's p-d intentions must be consistent with the agent's *ex ante* knowledge, as implied by the validity of $\Box K_\alpha\varphi \rightarrow I_\alpha\varphi$ and the validity of schema (D) for I_α imply that. In other words, if an agent p-d intends φ then the agent must not know $\neg\varphi$ *ex ante*. In light of the validity of $\Box K_\alpha\varphi \leftrightarrow K_\alpha\Box\varphi$ (a formula associated with frame condition ($\mathtt{Unif-H}$)), this implies that an agent cannot at the same time p-d intend φ and know that φ is impossible, since formula $I_\alpha\varphi \rightarrow \neg K_\alpha\Box\neg\varphi$ is valid.

[25] As for *intention-with-which*, a viable characterization can be conceived using an extension of *IEST* with a belief modality $B_\alpha\varphi$ (see Subsection 5.5.2), as follows: suppose that at a given index agent α had a primary reason ψ for choosing its current action, so that $I^d_\alpha\psi \wedge (\Box B_\alpha([\alpha]\varphi \rightarrow \psi) \wedge ([\alpha]\varphi \wedge I^d_\alpha[\alpha]\varphi))$ holds. Thus, at the index (a) α deliberatively p-d intended ψ, (b) α believed *ex ante* that $[\alpha]\varphi \rightarrow \psi$, and (c) α has deliberative-intentionally seen to it that φ. Then the realization of ψ can be thought of as an intention-with-which α has seen to it that φ.

Secondly, agents always have positive and negative introspection about their p-d intentions with respect to their *ex ante* knowledge, as implied by the validity of formulas $I_\alpha\varphi \to \Box K_\alpha I_\alpha\varphi$ and $\neg I_\alpha\varphi \to \Box K_\alpha\neg I_\alpha\varphi$.

5.4.2 Axiomatization

Just as done in all the previous chapters, in this subsection I introduce a proof system for the developed logic:

Definition 5.8 (Proof system for *IEST*). *Let Λ_I be the proof system defined by the following axioms and rules of inference:*

- (Axioms) *All classical tautologies from propositional logic; the* **S5** *schemata for* $\Box$, $[\alpha]$, *and* K_α*; the* **KD** *schemata for* I_α*; and the following schemata:*

$$\begin{array}{ll}
\Box\varphi \to [\alpha]\varphi & (SET) \\
\textit{For all } m \geq 1 \textit{ and pairwise different } \alpha_1, \dots, \alpha_m, & \\
\bigwedge_{1\leq k\leq m} \Diamond[\alpha_i]\varphi_i \to \Diamond\left(\bigwedge_{1\leq k\leq m}[\alpha_i]\varphi_i\right) & (IA) \\
K_\alpha\varphi \to [\alpha]\varphi & (OAC) \\
\Diamond K_\alpha\varphi \to K_\alpha\Diamond\varphi & (Unif-H) \\
\Box K_\alpha\varphi \to I_\alpha\varphi & (InN) \\
I_\alpha\varphi \to \Box K_\alpha I_\alpha\varphi & (KI)
\end{array}$$

- *(Rules of inference) Modus Ponens, Substitution, and Necessitation for all modal operators.*

Schemata (*SET*) and (*IA*) are standard in *BST*, and they were discussed in Chapter 2's Subsection 2.3.1. Schemata (*OAC*) and ($Unif - H$) were discussed in Chapter 4's Subsection 4.5.2.

Schema (*InN*)—where 'InN' stands for *intentional necessity*—characterizes syntactically what I called the *knowledge-to-intention property* (p. 228, see also item 2c in the list of *IEST*'s logic-based properties in Subsection 5.4.1).

Schema (*KI*)—where 'KI' stands for *knowledge of intention*—characterizes syntactically the *knowledge-of-intention property* (p. 228, see also item 3 in the list of *IEST*'s logic-based properties in Subsection 5.4.1), as well as frame condition (`KI`) (see Definition 5.4).

Observation 5.9. *Schemata* (4) *and* (5) *for* I_α, *as well as schema* ($\star$) $\neg I_\alpha\varphi \to \Box K_\alpha\neg I_\alpha\varphi$ *and schema* (*Den*) $\Diamond I_\alpha\varphi \to K_\alpha\langle I_\alpha\rangle\varphi$, *are important* Λ_I*-theorems, which can be shown to be* Λ_I*-provable according to the following arguments:*

(a) For schema ($\star$) $\neg I_\alpha\varphi \to \Box K_\alpha\neg I_\alpha\varphi$, a derivation is obtained as follows: let $(KI)'$ denote schema $I_\alpha\varphi \to K_\alpha\Box I_\alpha\varphi$. Observe that Λ_I-theorem $\Box K_\alpha\varphi \leftrightarrow K_\alpha\Box\varphi$ (which is indeed a Λ_I-theorem since it is logically equivalent to $(Unif - H)$, as shown in Observation 4.31 b) implies that (KI) in Λ_I is logically equivalent to $(KI)'$. Thus, a derivation of $\neg I_\alpha\varphi \to K_\alpha\Box\neg I_\alpha\varphi$ can be obtained by substituting $I_\alpha\varphi$ for $\odot^S_\alpha\varphi$ and $(KI)'$ for $(s.Cl)$ in the derivation of item c in the same Observation 4.31. Once again, Λ_I-theorem $\Box K_\alpha\varphi \leftrightarrow K_\alpha\Box\varphi$ then implies that $\neg I_\alpha\varphi \to \Box K_\alpha\neg I_\alpha\varphi$ is therefore also a Λ_I-theorem.

(b) Schema (4) for I_α follows straightforwardly from schema (KI) and schema (InN). Schema (5) for I_α follows straightforwardly from ($\star$) $\neg I_\alpha\varphi \to \Box K_\alpha\neg I_\alpha\varphi$ (item a above) and schema (InN).

(c) For schema (Den) $\Diamond I_\alpha\varphi \to K_\alpha\langle I_\alpha\rangle\varphi$, a derivation is obtained as follows, where 'c.p.' abbreviates 'contrapositive,' 'Nec.' abbreviates 'Necessitation,' and 'Subs.' abbreviates 'Substitution':

1. $\vdash_{\Lambda_I}$	$\neg I_\alpha\varphi \to \Box\neg I_\alpha\varphi$	($\star$), Subs. of (T) for K_α, modal & prop. logic
2. $\vdash_{\Lambda_I}$	$\Diamond I_\alpha\varphi \to I_\alpha\varphi$	C.p. of 1
3. $\vdash_{\Lambda_I}$	$I_\alpha\varphi \to K_\alpha I_\alpha\varphi$	(KI), Subs. of (T) for $\Box$, modal & prop. logic
4. $\vdash_{\Lambda_I}$	$K_\alpha I_\alpha\varphi \to K_\alpha\langle I_\alpha\rangle\varphi$	(D) for I_α, Nec. & Subs. of (K) for K_α
5. $\vdash_{\Lambda_I}$	$\Diamond I_\alpha\varphi \to K_\alpha\langle I_\alpha\rangle\varphi$	2, 3, 4, prop. logic.

As for metalogic properties of *IEST*, the soundness & completeness results for Λ_I are stated in the following theorem, whose proof is relegated to Appendix D:

Theorem 5.10 (Soundness & Completeness of Λ_I). *The proof system Λ_I is sound and complete with respect to the class of* iebt-*models.*

□

The proof of Theorem 5.10 is the main technical contribution of this chapter. As for soundness, the proof is standard. As for completeness, the proof is a two-step process. First, I introduce a Kripke semantics for the logic—entirely based on relations on sets of possible worlds. In such a semantics, the formulas of $\mathcal{L}_{\mathsf{I}}$ are evaluated on Kripke-*ies*-models (Definition D.14). I prove completeness of Λ_I with respect to the class of these structures, via the well-known technique of canonical models. Secondly, a truth-preserving correspondence between Kripke-*ies*-models and a sub-class of *iebt*-models is used for proving completeness with respect to *iebt*-models via completeness with respect to Kripke-*ies*-models. The truth-preserving correspondence implies associating a topological model to a Kripke model, such that both satisfy the same formulas at same indices. This is done via so-called Alexandrov spaces (Definition D.16), with a technique inspired by Özgün (2017) (see also Baltag, Bezhanishvili, Özgün, & Smets, 2015; Baltag et al., 2016).

5.5 Conclusion

I want to conclude this chapter with a discussion of three topics: (a) possibilities for future work in the stit-theoretic formalization of future-directed intentions, (b) a few aspects of the interplay between my notion of intentionality and the notion of p-1 belief that was introduced in Subsection 3.5.1 of Chapter 3's conclusion, and (c) a first proposal for formalizing the category of motivational responsibility (see the discussion on Broersen's three categories of responsibility in Chapter 1, p. 5).

5.5.1 An Account of Future-Directed Intentions

My semantics of intentionality shies away from modelling future-directed intentions. However, I believe that xstit theory (see Chapter 2's Subsection 2.3.3, p. 56), coupled with the strategic-ability modality $\langle\langle\alpha\rangle\rangle^s\varphi$ (see, for instance, Broersen et al., 2006a;Horty, 2001, Chapter 7), might aid in the construction of a framework that would account for an interesting relation between explicit future-directed intentions and intentional action, which would prove useful in a finer-grained characterization of motivational responsibility.

Definition 5.11 (Syntax for intentional xstit theory with strategies)**.** *Given a finite set Ags of agent names and a countable set of propositions P, the grammar of the formal language $\mathcal{L}_{\mathsf{SX}}$ is given by*

$$\varphi ::= p \mid \neg\varphi \mid \varphi \wedge \varphi \mid \Box\varphi \mid X\varphi \mid [\alpha]\varphi \mid K_\alpha \mid I_\alpha\varphi \mid \langle\langle\alpha\rangle\rangle^s\varphi,$$

where $p \in P$ and $\alpha \in Ags$.

In this language, $\Box\varphi$, $[\alpha]\varphi$, $K_\alpha\varphi$, and $I_\alpha\varphi$ have the same meaning as in Definition 5.3; $X\varphi$ expresses that 'φ holds at the next moment (along the same history)'; and $\langle\langle\alpha\rangle\rangle^s\varphi$ expresses that agent α has the strategic ability to ensure that φ is the case. As for the semantics, the formulas of $\mathcal{L}_{\mathsf{SX}}$ are evaluated on what I refer to as *uniformly bounded* iebdt-*models*, which are nothing more than finite *iebt*-models (see Definition 5.4) that additionally satisfy the following two conditions:

- (TD) *Time-discreteness*: for all $h \in H$ and $m \in h$ such that m is not terminal, there exists a unique moment m^{+h} such that $m \sqsubset m^{+h}$ and $m^{+h} \sqsubseteq m'$ for every $m' \in h$ such that $m \sqsubset m'$. For $m \in M$ and $h \in H_m$, m^{+h} is known as the *successor of m along h*. For $m \in M$, $h \in H_m$, the moment m^{+h} will also be denoted by $m^{+h(1)}$, the moment $\left(m^{+h}\right)^{+h}$ will also be denoted by $m^{+h(2)}$, so that, for $i \in \mathbb{N} - \{0\}$, $m^{+h(i)}$ will denote the unique moment in h that is the i^{th} iteration of the successor function applied to m. For the sake of coherence, $m^{+h(0)}$ will also denote m. For *index* $\langle m, h\rangle$, I refer to $\left\langle m^{+h}, h\right\rangle$ as the *successor of* $\langle m, h\rangle$ or as $\langle m, h\rangle$'s *next index*.

- (UB) *Uniform bound*: for all $h, h' \in H$, $card(h) = card(h')$.

For $\alpha \in Ags$ and moment m, let $\sqsubseteq [m] := \{m' \in M; m \sqsubseteq m'\}$. Then a *strategy of α starting at m* is defined as a function $\sigma : \sqsubseteq [m] \to \bigcup_{m' \in \sqsubseteq [m]} \mathbf{Choice}_{\alpha}^{m'}$ such that $\sigma(m') \in \mathbf{Choice}_{\alpha}^{m'}$. In other words, a strategy of α starting at m assigns to m' a choice of action available to α at m' (for every m' that is either equal to m or in the future of m). For $\alpha \in Ags$, moment m, and strategy σ_α starting at m, the set

$$\mathbf{Adm}_{\alpha}^{m}(\sigma_\alpha) := \left\{ h' \in H_m; \begin{array}{l} h' \in \sigma_\alpha(m') \\ \text{for every } m' \in h' \text{ s. t. } m' \sqsupseteq m \end{array} \right\}$$

is known as the set of *admissible* histories of σ_α. The idea is that strategy σ_α constrains the possible courses of events at m to the histories in $\mathbf{Adm}_{\alpha}^{m}(\sigma_\alpha)$.

To clarify, strategies underlie alternatives for sequential actions over time. Thus, they allow us to reason about the effects of chains of actions in the long run. If α is carrying out a particular strategy σ_α starting at m, then α will perform the action $\sigma_\alpha(m')$ recommended by that strategy whenever α arrives at m'. In this way, α sequentially constrains the possible futures while carrying out σ_α. Any history that results from this process of sequential constraining is said to be admitted by σ_α—and lies in the set $\mathbf{Adm}_{\alpha}^{m}(\sigma_\alpha)$.[26] With such a notion, one can define semantics for the strategic-ability modality:

Definition 5.12. *Let $\mathcal{M}$ be a uniformly bounded* iebdt-*model. The semantics on $\mathcal{M}$ for the formulas of $\mathcal{L}_{SX}$ are recursively defined as in Definition 5.5, with the following additional clauses:*

$$\begin{array}{lll} \mathcal{M}, \langle m, h\rangle \models X\varphi & \textit{iff} & \mathcal{M}, \langle m^{+h}, h\rangle \models \varphi \\ \mathcal{M}, \langle m, h\rangle \models \langle\langle \alpha \rangle\rangle^{s}\varphi & \textit{iff} & \textit{there is a strategy } \sigma_\alpha \textit{ starting at } m \textit{ s. t.} \\ & & \mathcal{M}, \langle m, h'\rangle \models \varphi \textit{ for every } h' \in \mathbf{Adm}_{\alpha}^{m}(\sigma_\alpha). \end{array}$$

[26] This concept of 'strategy' was first incorporated into stit theory by Belnap et al. (2001) and by Horty (2001). Similar to what is defined in extensive-form games or in concurrent game structures for alternating-time temporal logic (*ATL*), an agent's strategies are functions that map a given moment to a choice of action available to the agent at that moment.

Now, as mentioned on p. 220, I follow Bratman (1984, 1987) in considering future-directed intentions as elements in complex plans. In my view, the notion of strategies introduced above—which underlies modality $\langle\langle\alpha\rangle\rangle^s\varphi$—can be added to my theory of intentionality to formalize having plans, and thus to formalize a version of future-directed intentions. The idea is that an agent α has a future-directed intention—written f-d intention, from here on—toward the final realization of φ iff (a) there is a strategy by which α can enforce φ in the end, and (b) at each step of this strategy α p-d intends that φ will hold in the end. The sequence of actions recommended by the strategy, each coupled with the respective moment's p-d intention that φ will hold in the end, can be identified with a *plan* of α toward the final realization of φ. Thus, α f-d intends φ iff α plans that φ will hold in the end.

To express these ideas in all formality, for a non-terminal moment m and $h \in H_m$, let $i(m)$ denote the number of moments between m and h's terminal moment (including h's terminal moment but not including m).[27] Then I will say that at $\langle m,h\rangle$ α f-d intended φ iff $\mathcal{M},\langle m,h\rangle \models \langle\langle\alpha\rangle\rangle^s X^{i(m)}\varphi$ via strategy σ_α and, for all $h' \in \mathbf{Adm}^m_\alpha(\sigma_\alpha)$ and $0 \leq k < i(m)$, $\mathcal{M},\left\langle m^{+h'(k)},h'\right\rangle \models I_\alpha X^{i(m)-k}\varphi$ via p-d intention $U^{h'}_k$.[28] Thus, at $\langle m,h\rangle$ α f-d intended φ iff (a) there is a strategy starting at m by which α can enforce that φ holds at all indices based on terminal moments and anchored by histories admitted by the strategy, and (b) at all current and successor indices that are anchored by strategically admitted histories, α p-d intended that φ would hold in the end. The sequence of pairs $\left\{\left(\sigma_\alpha\left(m^{+h'(k)}\right),U^{h'}_k\right)\right\}_{h'\in\mathbf{Adm}^m_\alpha(\sigma_\alpha),0\leq k<i(m)}$ can be thought of as α's plan toward the final realization of φ.

To illustrate this semantics, consider the uniformly bounded *iebdt*-model depicted in Figure 5.6. In this simple example, agent *driver* intends to start her car. At moment m_1 there are two available choices: L_1, standing for the choice of getting the keys of the car, and L_2, standing for the choice of not getting the keys of the car. At m_2 the available choices are: L_3, standing for the choice of turning the key in the ignition, and L_4, standing for the choice of not turning the key in the ignition. At m_3 the available choices (L_5 and L_6) are irrelevant.

Focusing on *driver*'s intentionality, let $\tau^{\langle m_1,h_1\rangle}_{driver} = \left\{\emptyset,\pi^{\Box}_{driver}[\langle m_1,h_1\rangle],\{\langle m_1,h_1\rangle\}\right\}$, let $\tau^{\langle m_2,h_1\rangle}_{driver} = \left\{\emptyset,\pi^{\Box}_{driver}[\langle m_2,h_1\rangle],\{\langle m_2,h_1\rangle\}\right\}$, and let $\tau^{\langle m_3,h_3\rangle}_{driver} = \left\{\emptyset,\pi^{\Box}_{driver}[\langle m_3,h_3\rangle]\right\}$. These topologies are represented in the diagram using circles, where at m_1 the

[27] In other words $i(m) = card(\sqsubset[m] \cap h)$.

[28] Recall that $X^n\psi$ ($n \in \mathbb{N}$) denotes the formula that results from applying n-iterations of operator X behind φ (see Footnote 32 in Chapter 3, p. 104).

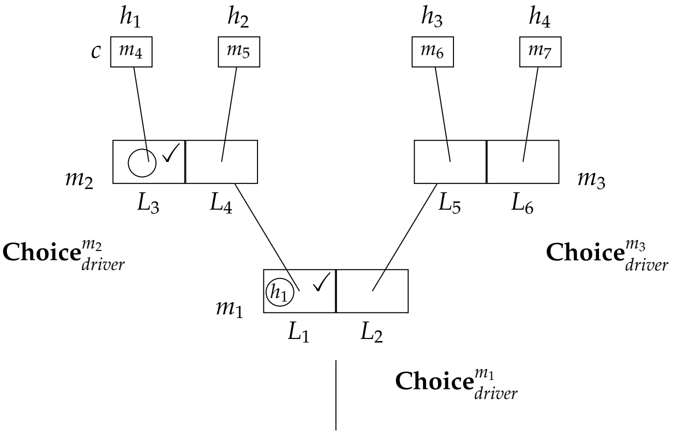

h_1
h_2
h_3
h_4
c
m_4
m_5
m_6
m_7
✓
m_2
m_3
L_3
L_4
L_5
L_6
$\textbf{Choice}^{m_2}_{driver}$
$\textbf{Choice}^{m_3}_{driver}$
h_1
✓
m_1
L_1
L_2
$\textbf{Choice}^{m_1}_{driver}$

Figure 5.6: *Another driver example.*

circle encloses the tag 'h_1' to rule out a p-d intention including $\langle m_1, h_2 \rangle$, and where for readability I omitted displaying ellipses that would represent full *ex ante* information sets.

Let σ_{driver} be a strategy starting at m_1 such that $\sigma_{driver}(m_1) = L_1$, $\sigma_{driver}(m_2) = L_3$ (where these recommendations are represented in the diagram using checkmarks), and $\sigma_{driver}(m)$ is arbitrary for every other moment m lying in m_1's future. Thus, $\textbf{Adm}^{m_1}_{driver}(\sigma_{driver}) = \{h_1\}$. Let c stand for the proposition '*driver*'s car is started.' Then the diagram implies that $\mathcal{M}, \langle m_4, h_1 \rangle \models c$: *at* $\langle m_4, h_1 \rangle$ driver*'s car has started.* Now, observe that σ_{driver} implies that $\mathcal{M}, \langle m_1, h_1 \rangle \models \langle\langle driver \rangle\rangle^s X^2 c$: *at* $\langle m_1, h_1 \rangle$ driver *had the strategic ability to start her car in the end—by first getting the key and then turning it in the ignition.* Furthermore, observe that both $\mathcal{M}, \langle m_1, h_1 \rangle \models I_{driver} X^2 c$ (via $U_1 := \{\langle m_1, h_1 \rangle\}$) and $\mathcal{M}, \langle m_2, h_1 \rangle \models I_{driver} X c$ (via $U_2 := \{\langle m_2, h_1 \rangle\}$): *at both* $\langle m_1, h_1 \rangle$ *and* $\langle m_2, h_1 \rangle$ driver *p-d intended that her car would start in the end.* Since $i(m_1) = 2$, $\{(L_1, U_1), (L_3, U_2)\}$ is a plan of *driver* toward the final realization of the starting of her car, and one can say that at $\langle m_1, h_1 \rangle$ *driver* f-d intended that her car would start.

This version of future-directed intentions can be used to characterize intentional action so that the relation between the two notions does not fall under the *Simple View* (see item 1c in the list of *IEST*'s logic-based properties in Subsection 5.4.1, as well as Footnote 22). Inspired by Bratman's (1984) idea of *motivational potential,* let me characterize intentional action as follows: at $\langle m, h \rangle$ α has intentionally seen to it that φ iff (a) at $\langle m, h \rangle$ α f-d intended ψ via plan $\left\{\left(\sigma_\alpha\left(m^{+h'(k)}\right), U^{h'}_k\right)\right\}_{h' \in \textbf{Adm}^m_\alpha(\sigma_\alpha), 0 \leq k < i(m)}$, (b) $h \in \sigma_\alpha(m)$, and (c) at $\langle m, h \rangle$ α knew that

φ is an effect of $\sigma_\alpha(m)$ ($K_\alpha[\alpha]\varphi$ holds at $\langle m,h\rangle$). In other words, at an index an agent has intentionally seen to it that φ iff at the index (a) the agent had an f-d intention toward the final realization of ψ, (b) the agent chose the action currently recommended by its strategy (toward the final realization of the agent's f-d intention), and (c) the agent knew that φ is an effect of such a recommendation.[29]

This characterization does not enforce that intentionally doing must imply f-d intending. In the example above, for instance, at $\langle m_1,h_1\rangle$ *driver* has intentionally grabbed the keys, and at $\langle m_2,h_1\rangle$ *driver* has intentionally turned the key in the ignition—where both these intentional actions were carried out with the intention that the car would be started. Depending on the valuation of atomic propositions standing for 'grabbing the keys' and 'turning the key in the ignition,' one can both render models where *driver* f-d intended to *have* grabbed the keys and f-d intended to *have* turned the key in the ignition, on the one hand, and render models where *driver* did not f-d intend to grab the keys and did not f-d intend to turn the key in the ignition, on the other. Models of the latter kind could help in formalizing a potential solution to Bratman's (1984) target example (see Footnote 22).

Now, these stit-theoretic versions of f-d intentions and of intentional action are part of a merely initial proposal—which might not seem entirely satisfactory to some. Tailoring this proposal, as well as exploring its logic, is left for future work.

5.5.2 Intentionality & P-1 Belief

Many of the works reviewed in this chapter presuppose that intentionality is deeply connected with belief (Bratman, 1984; Cohen & Levesque, 1990; Herzig & Longin, 2004; Rao & Georgeff, 1991; Wooldridge, 2000). Moreover, and as often mentioned in this thesis, I consider that belief is an important epistemic component of responsibility. This is why the conclusions of Chapters 3 and 4 briefly explored particular extensions of *EST* with probabilistic belief, in the context of building nuanced theories of responsibility that would include an account of belief-driven choice. Once again leaving full-fledged analyses for future work, here I carry on the discussions of those chapters' conclusions, with the goal of sketching out interesting aspects of a potential merged framework. To be precise, I address three properties concerning the relation between this chapter's intentionality and the notion of p-1 belief from Chapter 3's conclusion (Subsection 3.5.1):

[29] Observe that, in contrast to the version of intentional action that was addressed in the other sections of this chapter, this version of intentional action does adhere to what Marcus (2019) called the *knowledge thesis* of intentional action, according to which an agent intentionally does φ only if the agent knowingly does φ (see item 2b in the list of *IEST*'s logic-based properties in Subsection 5.4.1).

- $\not\models I_\alpha\varphi \rightarrow B_\alpha\Diamond\varphi$: it is not necessarily true that if at an index an agent p-d intended φ then the agent must have p-1 believed that φ was currently possible. An example can be obtained from the anesthesiologist example depicted in Figure 5.2, by defining μ_{doctor} and $\tau^{\langle m_2,h_1\rangle}_{doctor}$ according to the following story: suppose that *doctor* intended to kill the patient by supplying anesthetics on a full stomach, but at the moment reserved for supplying anesthetics she p-1 believed that the patient had an empty stomach. Thus, let μ_{doctor} be defined so that $\mu_{doctor}\left(\{\langle m_2,h_1\rangle\} \mid \pi_{doctor}\left[\langle m_2,h_1\rangle\right]\right) = 1$ and $\mu_{doctor}\left(\{\langle m_2,h_2\rangle\} \mid \pi_{doctor}\left[\langle m_2,h_2\rangle\right]\right) = 1$, and let $\tau^{\langle m_2,h_1\rangle}_{doctor} = \left\{\emptyset, \pi^{\Box}_{doctor}\left[\langle m_2,h_1\rangle\right], \{\langle m_3,h_3\rangle\}\right\}$. This implies that $\mathcal{M},\langle m_2,h_1\rangle \models I_{doctor}[doctor]d \wedge \neg B_{doctor}\Diamond d$: *at* $\langle m_2,h_1\rangle$ doctor *p-d intended to kill the patient, but* doctor *did not p-1 believe that killing the patient was possible.*

- $\not\models I_\alpha\varphi \rightarrow \neg B_\alpha\neg\varphi$: it is not necessarily true that if at an index an agent p-d intended φ then the agent must have not p-1 believed $\neg\varphi$. An example can be found in the previous item's scenario. Observe that the definition of μ_{doctor} implies that $\mathcal{M},\langle m_2,h_1\rangle \models I_{doctor}[doctor]d \wedge B_{doctor}\neg d$: *at* $\langle m_2,h_1\rangle$ doctor *p-d intended to kill the patient, but* doctor *believed that the patient would live.* In fact, $\mathcal{M},\langle m_2,h_1\rangle \models B_{doctor}\Box\neg d$: *at* $\langle m_2,h_1\rangle$ doctor *p-1 believed that to kill the patient was impossible.*

- $\not\models (I_\alpha\varphi \wedge B_\alpha\Box(\varphi \rightarrow \psi)) \rightarrow I_\alpha\psi$: p-d intentions are not closed under (*ex ante*) belief of side-effect implication. An example can be obtained from the previous items' scenario. Observe that $\mathcal{M},\langle m_2,h_1\rangle \models (I_{doctor}a \wedge B_{doctor}\Box(a \rightarrow r)) \wedge \neg I_{doctor}r$: *at* $\langle m_2,h_1\rangle$ doctor *p-d intended that the anesthetics were supplied and p-1 believed that it was settled that supplying anesthetics implied that the patient would get ready for surgery; still,* doctor *did not intend that the patient would get ready for surgery.*

The first two properties imply that the potential merged framework would deviate from assumptions that are common to most frameworks formalizing both intentions and beliefs (Cohen & Levesque, 1990; Herzig & Schwarzentruber, 2008; Konolige & Pollack, 1993; Lorini & Herzig, 2008). The justification for not enforcing these assumptions comes from my interpretation of the semantics for modalities $B_\alpha\varphi$ and $I_\alpha\varphi$. On the one hand, p-1 belief refers to a notion known as 'certain belief' (Baltag & Smets, 2008), so that an agent could have had a p-d intention toward the realization of φ and not have certainty that φ is possible. On the other hand, the instantaneous nature of p-d intentions justify the idea that an agent could have had a p-d intention of φ and at the same moment have come to p-1 believe $\neg\varphi$, just then.

The third property, in turn, implies that the potential merged framework would solve the most usual formulations of the side-effect problem (see Footnotes 8 and 19).

5.5.3 A Glimpse at Motivational Responsibility

As demonstrated in Subsection 3.5.2 of Chapter 3's conclusion, one can formalize the categories of causal and informational responsibility—the first two categories in Broersen's classification—using the models for knowledge and agency given in previous chapters. The contents of this chapter, then, aid in the formalization of the third category: motivational responsibility. Recall from Chapter 1 (p. 5) that an agent is *causally responsible* for a state of affairs iff the agent is the material author of such a state. In turn, an agent is *motivationally responsible* for a state of affairs iff the agent is the material author and it behaved knowingly and intentionally while bringing about the state of affairs. Thus, in Example 5.6, for instance, *driver* should be held both causally responsible and motivationally responsible for killing the traffic officer. In Example 5.7, in contrast, although *doctor* should be held casually responsible for killing the patient, she should not be held motivationally responsible.

The discussion in Subsection 3.5.2 of Chapter 3's conclusion implies that one can think of formula $[\alpha]\varphi \wedge \Diamond\neg[\alpha]\varphi$ as a good candidate for syntactically characterizing causal responsibility (see also Lorini et al., 2014): agent α was causally responsible for φ iff α has seen to it that φ and it was possible for α to refrain from seeing to it that φ. In turn, formula $K_\alpha[\alpha]\varphi \wedge I_\alpha[\alpha]\varphi \wedge K_\alpha\Diamond\neg[\alpha]\varphi$ is a likewise good candidate for syntactically characterizing motivational responsibility: α was motivationally responsible for φ iff α has knowingly and intentionally seen to it that φ and α knew that it was possible to refrain from seeing to it that φ. This follows the ideas behind the notion of deliberative intentional action discussed in Subsection 5.3.1 (p. 228), so that an agent will be motivationally responsible for φ only if it deliberatively intended to bring about φ, and furthermore the agent knew that it was possible to refrain. Otherwise, someone could claim that the agent's not knowing that it was possible to refrain was what led it to intend to see to it that φ and to knowingly do so, so that the agent was not really motivationally responsible for φ.[30] To illustrate this characterization, consider Example 5.6. Observe that $\mathcal{M}, \langle m_1, h_1\rangle \models K_{driver}[driver]k \wedge I_{driver}[driver]k \wedge K_{driver}\Diamond\neg[driver]k$: *at* $\langle m_1, h_1\rangle$ driver

[30] Indeed, in the case of blameworthy agents, claims of this kind amount to excuses, and this is why in Chapter 6's Subsection 6.3.2 I refer to the policy that allows such excuses as *lenient on blameworthy agents*.

was motivationally responsible for killing the traffic officer. In turn, in Example 5.7 $\mathcal{M}, \langle m_3, h_3 \rangle \models [doctor]d \land \neg I_{doctor}[doctor]d \land K_{doctor}\Diamond\neg[doctor]d$: *at* $\langle m_3, h_3 \rangle$ doctor *was causally responsible for the patient's death, but she was not motivationally responsible.*

As it turns out, these candidate-formulas greatly influenced Chapter 6's (Subsection 6.3.2) proposal for the syntactic characterization of causal, resp. motivational, responsibility. Thus, on the road to formalizing Broersen's categories of responsibility, we are already equipped with tools to characterize all three of them.

Appendix D Metalogic Results for *IEST*

D.1 Soundness

Proposition D.13 (Soundness of Λ_I). *The system Λ_I is sound with respect to the class of* iebt-*models.*

Proof. The proof of soundness is routine: the validity of the **S5** schemata for $\Box$ and $[\alpha]$, as well as that of (*SET*) and (*IA*), is standard from *BST*; the validity of the **S5** schemata for K_α is standard from *EST*; the validity of (*OAC*) and ($Unif-H$) is shown exactly as in Chapter 4's Proposition C.40; the validity of the **KD** schemata for I_α (*InN*) follows from Definitions 5.4 and 5.5; and the validity of (*KI*) follows from frame condition (KI). □

D.2 Completeness

To prove completeness of Λ_I with respect to *iebt*-models, we will first prove completeness with respect to a class of Kripke models. The reason is that there exists a truth-preserving correspondence between this class and a sub-class of *iebt*-models. Below, I define said class of Kripke models and prove such a truth-preserving correspondence.

Definition D.14 (Kripke-*ies*-frames & models). *A tuple*

$$\left\langle W, Ags, R_\Box, Ags, \mathtt{Choice}, \{\approx_\alpha\}_{\alpha\in Ags}, \left\{R^I_\alpha\right\}_{\alpha\in Ags}\right\rangle$$

*is called a Kripke-*ies*-frame (where the acronym 'ies' stands for 'epistemic intentional stit') iff*

- *$W, Ags, R_\Box, Ags, \mathtt{Choice}$, and $\{\approx_\alpha\}_{\alpha\in Ags}$ are defined exactly as in Definition C.41 (p. 200).*

  For $\alpha \in Ags$ and $w \in W$, α's ex ante information set *at w is defined as $\pi^\Box_\alpha[w] := \{v; w \approx_\alpha \circ R_\Box v\}$, which by frame condition $(\mathtt{Unif}-\mathtt{H})_\mathtt{K}$ coincides with the set $\{v; wR_\Box \circ \approx_\alpha v\}$. To clarify, $(\mathtt{Unif}-\mathtt{H})_\mathtt{K}$ implies that $R_\Box \circ \approx_\alpha = \approx_\alpha \circ R_\Box$. Thus, $\approx_\alpha \circ R_\Box$ is an equivalence relation such that $\pi^\Box_\alpha[w] = \pi^\Box_\alpha[v]$ for every $w, v \in W$ such that $w \approx_\alpha \circ R_\Box v$.*

- *For $\alpha \in Ags$, R^I_α is a serial, transitive, and euclidean relation on W such that $R^I_\alpha \subseteq \approx_\alpha \circ R_\Box$ and such that the following condition is satisfied:*

  - $(\mathtt{Den})_\mathtt{K}$ *For all $v, u \in W$ such that $v \approx_\alpha \circ R_\Box u$, there exists $z \in W$ such that $vR^I_\alpha z$ and $uR^I_\alpha z$.*

> *For $\alpha \in Ags$, R^{I+}_{α} denotes the reflexive closure of R^{I}_{α}. For $w \in W$, $w \uparrow_{R^{I+}_{\alpha}}$ denotes the set $\left\{v \in W; wR^{I+}_{\alpha}v\right\}$.*

*A Kripke-*ies*-model $\mathcal{M}$ consists of the tuple that results from adding a valuation function $\mathcal{V}$ to a Kripke-*ies*-frame, where $\mathcal{V} : P \to 2^{W}$ assigns to each atomic proposition a set of worlds (recall that P is the set of propositions in $\mathcal{L}_I$).*

Kripke-*ies*-models allow us to evaluate the formulas of $\mathcal{L}_\mathsf{I}$ with semantics that are analogous to the ones provided for *iebt*-frames:

Definition D.15 (Evaluation rules on Kripke models)**.** *Let $\mathcal{M}$ be a Kripke-*ies*-model. The semantics on $\mathcal{M}$ for the formulas of $\mathcal{L}_I$ are defined recursively by the following truth conditions, evaluated at world w:*

$$\begin{array}{lll}
\mathcal{M}, w \models p & \textit{iff} & w \in \mathcal{V}(p) \\
\mathcal{M}, w \models \neg\varphi & \textit{iff} & \mathcal{M}, w \not\models \varphi \\
\mathcal{M}, w \models \varphi \wedge \psi & \textit{iff} & \mathcal{M}, w \models \varphi \textit{ and } \mathcal{M}, w \models \psi \\
\mathcal{M}, w \models \Box\varphi & \textit{iff} & \textit{for all } v \in \overline{w}, \mathcal{M}, v \models \varphi \\
\mathcal{M}, w \models [\alpha]\varphi & \textit{iff} & \textit{for all } v \in \mathtt{Choice}^{\overline{w}}_{\alpha}(w), \mathcal{M}, v \models \varphi \\
\mathcal{M}, w \models K_{\alpha}\varphi & \textit{iff} & \textit{for all } v \textit{ s. t. } w \approx_{\alpha} v, \mathcal{M}, v \models \varphi \\
\mathcal{M}, w \models I_{\alpha}\varphi & \textit{iff} & \textit{there exists } x \in \pi^{\Box}_{\alpha}[w] \textit{ s. t. } x \uparrow_{R^{I+}_{\alpha}} \subseteq |\varphi|,
\end{array}$$

where I write $|\varphi|$ to refer to the set $\{w \in W; \mathcal{M}, w \models \varphi\}$. Satisfiability, validity, and general validity are defined as usual.

Importantly, Kripke-*ies*-models can be used for constructing *iebt*-models such that both satisfy the same formulas of $\mathcal{L}_I$. Such a construction implies defining topologies on the basis of relations, so that the following definition and observation are very important.

Definition D.16 (Alexandrov spaces)**.** *A topological space (X, τ) is said to be an* Alexandrov space *iff the intersection of any collection of open sets of X is an open set as well.*

Notice that a space is Alexandrov iff every point $x \in X$ has a $\subseteq$-smallest open set including it, namely the intersection of all the open sets around x.

Definition D.17. *For a given frame (X, R) such that R is reflexive and transitive, a set $A \subseteq X$ is called upward-closed iff for all $x \in A$, if $x \leq y$ for some $y \in X$, then $y \in A$ as well. For $x \in X$, $x \uparrow_R$ denotes the set $\{y \in X; xRy\}$, which is clearly upward closed.*

Observation D.18. *For every frame (X, R) such that R is reflexive and transitive, the set of all R-upward-closed sets forms an Alexandrov topology on X, denoted by τ_R. For all $x \in X$, the $\subseteq$-smallest open set including x is precisely $x \uparrow_R$. This implies that $\{x \uparrow_R ; x \in X\}$ is a basis for the topology τ_R.*

Proof. It is routine. The reader is referred to van Benthem and Bezhanishvili (2007, Section 2) for details (see also Özgün, 2017, Chapter 3). □

We are now ready to define, for a Kripke-*ies*-frame, an associated *iebt*-frame.

Definition D.19 (Associated *iebt*-frame)**.** *Let*

$$\mathcal{F} = \left\langle W, Ags, R_\Box, \texttt{Choice}, \{\approx_\alpha\}_{\alpha \in Ags}, \left\{R^I_\alpha\right\}_{\alpha \in Ags} \right\rangle$$

*be a Kripke-*ies*-frame. Then $\mathcal{F}^T := \left\langle M_W, \sqsubset, Ags, \mathbf{Choice}, \{\sim_\alpha\}_{\alpha \in Ags}, \tau \right\rangle$ is called the* iebt*-frame associated with $\mathcal{F}$ iff*

- *$M_W, \sqsubset, \mathbf{Choice}$, and $\{\sim_\alpha\}_{\alpha \in Ags}$ are defined just as in Chapter 4's Definition C.43 (p. 201).*
- *τ is a function defined as follows:*
  - *For $\alpha \in Ags$ and $z \in W$, $\tau_\alpha^{\langle z, h_z \rangle} = \{\emptyset, \pi_\alpha^\Box [\langle z, h_z \rangle]\}$.*
  - *For $\alpha \in Ags$, we first define a relation R^{IT}_α on $\{\langle \overline{w}, h_v \rangle ; w \in W \text{ and } v \in \overline{w}\}$ by the rule: $\langle \overline{w}, h_v \rangle R^{IT}_\alpha \left\langle \overline{w'}, h_{v'} \right\rangle$ iff $vR^I_\alpha v'$.*

    For $\alpha \in Ags$, $w \in W$, and $v \in \overline{w}$, then, we define $\tau_\alpha^{\langle \overline{w}, h_v \rangle}$ as the subspace topology of $\tau_{R^{IT+}_\alpha}$ (the Alexandrov topology induced by relation R^{IT+}_α according to Observation D.18) on $\pi_\alpha^\Box [\langle \overline{w}, h_v \rangle]$.[31] Observe that, for all $\alpha \in Ags$, $w \in W$, and $v \in \overline{w}$, $\pi_\alpha^\Box [\langle \overline{w}, h_v \rangle] = \left\{\left\langle \overline{v'}, h_{v'} \right\rangle ; v' \in \pi_\alpha^\Box[v]\right\}$. Thus, the fact that $R^I_\alpha \subseteq \approx_\alpha \circ R_\Box$ implies that, for all $\langle \overline{x}, h_x \rangle \in \pi_\alpha^\Box [\langle \overline{w}, h_v \rangle]$, $\langle \overline{x}, h_x \rangle \uparrow_{R^{IT+}_\alpha} \subseteq \pi_\alpha^\Box [\langle \overline{w}, h_v \rangle]$, so that $\pi_\alpha^\Box [\langle \overline{w}, h_v \rangle]$ is open in $\tau_{R^{IT+}_\alpha}$.
  - *For $\alpha \in Ags$ and $v \in W$, $\tau_\alpha^{\langle W, h_v \rangle} = \{\emptyset, \pi_\alpha^\Box [\langle W, h_v \rangle]\}$.*

Proposition D.20. *Let $\mathcal{F}$ be a Kripke-*ies*-frame. Then $\mathcal{F}^T$ is an* iebt*-frame, indeed.*

Proof. It amounts to showing that $\sqsubset$ is a strict partial order that satisfies no backward branching, that **Choice** is a function that satisfies frame conditions (`NC`) and (`IA`), that $\{\approx_\alpha\}_{\alpha \in Ags}$ is such that $\approx_\alpha$ is an equivalence relation for every $\alpha \in Ags$ and frame conditions (`OAC`) and (`Unif` − `H`) are met, and that τ is a function that meets the requirements of Definition 5.4.

[31] Let τ be a topology on X. For each $A \subseteq X$, the subspace topology of τ on A is the family $\{U \cap A \,| U \in \tau\}$.

- For all the properties mentioned above, except the one concerning τ, the proofs are exactly the same as their analogs' in Chapter 4's Proposition C.44 (p. 202).
- As for τ let us show that it satisfies conditions (CI) and (KI).

  Observe that, for each $\alpha \in Ags$ and each index $\langle m, h\rangle$ either of the form $\langle z, h_z\rangle$ (where $z \in W$) or of the form $\langle W, h_v\rangle$ (where $v \in W$), $\tau_\alpha^{\langle m,h\rangle}$ is defined as topology on $\pi_\alpha^\Box[\langle m, h\rangle]$ that trivially satisfies the conditions imposed both by (CI) and by (KI). Thus, τ satisfies (CI) and (KI) at such indices.

  Assume, then, that $\langle m, h\rangle$ is of the form $\langle \overline{w}, h_v\rangle$, where $v \in \overline{w}$. Take $\alpha \in Ags$. By Definition D.19, $\tau_\alpha^{\langle \overline{w},h_v\rangle}$ is the subspace topology of $\tau_{R_\alpha^{IT+}}$ on $\pi_\alpha^\Box[\langle \overline{w}, h_v\rangle]$. Thus, $\tau_\alpha^{\langle \overline{w},h_v\rangle}$ is a topology on $\pi_\alpha^\Box[\langle \overline{w}, h_v\rangle]$ that by definition satisfies the condition imposed by (KI). Let us show that the condition imposed by (CI) is also satisfied. Take $U, V \in \tau_\alpha^{\langle \overline{w},h_v\rangle}$ such that U and V are non-empty. Take $\langle \overline{u}, h_u\rangle \in U$ and $\langle \overline{x}, h_x\rangle \in V$. Definition D.19—and in turn Definition C.43 (p. 201)—implies that $u \approx_\alpha \circ R_\Box x$. $\mathcal{F}$'s condition $(\mathtt{Den})_\mathtt{K}$ implies that there exists $z \in W$ such that $uR_\alpha^I z$ and $xR_\alpha^I z$, which by definition of R_α^{IT} implies that $\langle \overline{u}, h_u\rangle R_\alpha^{IT} \langle \overline{z}, h_z\rangle$ and that $\langle \overline{x}, h_x\rangle R_\alpha^{IT} \langle \overline{z}, h_z\rangle$. Thus, $\langle \overline{z}, h_z\rangle \in \langle \overline{u}, h_u\rangle \uparrow_{R_\alpha^{IT+}}$ and $\langle \overline{z}, h_z\rangle \in \langle \overline{x}, h_x\rangle \uparrow_{R_\alpha^{IT+}}$. Since $\pi_\alpha^\Box[\langle \overline{w}, h_v\rangle]$ is open in $\tau_{R_\alpha^{IT+}}$ (see the second bullet point in Definition D.19), we know that $\langle \overline{u}, h_u\rangle \uparrow_{R_\alpha^{IT+}} \subseteq U$ and that $\langle \overline{x}, h_x\rangle \uparrow_{R_\alpha^{IT+}} \subseteq V$. Thus, $\langle \overline{z}, h_z\rangle \in U \cap V$, so that U (and V) is $\tau_\alpha^{\langle \overline{w},h_v\rangle}$-dense.

□

Let $\mathcal{M}$ be a Kripke-*ies*-model with valuation function $\mathcal{V}$. The frame upon which $\mathcal{M}$ is based has an associated *iebt*-frame. If to the tuple of this *iebt*-frame one adds a valuation function $\mathcal{V}^T$ such that $\mathcal{V}^T(p) = \{\langle \overline{w}, h_w\rangle\, ; w \in \mathcal{V}(p)\}$, the resulting model is called the *iebt*-model associated with $\mathcal{M}$.

Proposition D.21 (Truth-preserving correspondence)**.** *Let $\mathcal{M}$ be a Kripke-*ies*-model, and let $\mathcal{M}^T$ denote its associated *iebt*-model. For all φ of $\mathcal{L}_I$ and $w \in W$, $\mathcal{M}, w \models \varphi$ iff $\mathcal{M}^T, \langle \overline{w}, h_w\rangle \models \varphi$.*

Proof. We proceed by induction on the complexity of φ. For the base case, the cases of Boolean connectives, and the cases of all modal operators except I_α, the proofs are exactly the same as their analogs' in Chapter 4's Proposition C.46 (p. 204). As for the case of I_α, the following arguments complete the induction proof:

- ("I_α") First, observe that, by induction hypothesis, $\|\varphi\| = \{\langle \overline{w}, h_w\rangle\, ; w \in |\varphi|\}$.

Therefore, $\mathcal{M}, w \models I_\alpha\varphi$ iff there exists $x \in \pi_\alpha^{\Box}[w]$ such that $x \uparrow_{R_\alpha^{I+}} \subseteq |\varphi|$ iff $\langle \overline{x}, h_x\rangle \uparrow_{R_\alpha^{IT+}} \subseteq \|\varphi\|$ iff there exists $U \in \tau_\alpha^{\langle \overline{w}, h_w\rangle}$ such that $U \subseteq \|\varphi\|$ iff $\mathcal{M}^T, \langle \overline{w}, h_w\rangle \models I_\alpha\varphi$.

□

Proposition D.21 implies that to prove completeness of Λ_I with respect to *iebt*-models all we need to do is prove completeness with respect to Kripke-*ies*-models. Therefore, let us prove completeness with respect to Kripke-*ies*-models, via the well-known technique of canonical models.

D.2.1 Canonical Kripke-*Ies*-Structure

We show that the proof system Λ_I is complete with respect to the class of Kripke-*ies*-models. For each Λ_I-consistent formula φ, we build a canonical structure from the syntax that satisfies φ.

Definition D.22 (Canonical Structure)**.** *The tuple*

$$\mathcal{M} = \left\langle W^{\Lambda_I}, R_\Box, \texttt{Choice}, \{\approx_\alpha\}_{\alpha\in Ags}, \left\{R_\alpha^I\right\}_{\alpha\in Ags}, \mathcal{V}\right\rangle$$

is called a canonical structure for Λ_I iff

- $W^{\Lambda_I} = \{w; w \text{ is a } \Lambda_I\text{-MCS}\}$. *$R_\Box$ is a relation over W^{Λ_I} defined by the rule: $wR_\Box v$ iff $\Box\varphi \in w \Rightarrow \varphi \in v$ for every φ of $\mathcal{L}_I$. For $w \in W^{\Lambda_I}$, the set $\left\{v \in W^{\Lambda_I}; wR_\Box v\right\}$ is denoted by $\overline{w}$.*
- $\texttt{Choice}$ *is a function that assigns to each α and $\overline{w}$ a subset of $2^{\overline{w}}$, denoted by $\texttt{Choice}_\alpha^{\overline{w}}$, and defined as follows: let $R_\alpha^{\overline{w}}$ be a relation on $\overline{w}$ such that, for $w, v \in W^{\Lambda_I}$, $wR_\alpha^{\overline{w}}v$ iff $[\alpha]\varphi \in w \Rightarrow \varphi \in v$ for every φ of $\mathcal{L}_I$; if $\texttt{Choice}_\alpha^{\overline{w}}(v) := \left\{u \in \overline{w}; vR_\alpha^{\overline{w}}u\right\}$, then $\texttt{Choice}_\alpha^{\overline{w}} := \left\{\texttt{Choice}_\alpha^{\overline{w}}(v); v \in \overline{w}\right\}$.*
- *For $\alpha \in Ags$, $\approx_\alpha$ is an epistemic relation on W^{Λ_I} given by the rule: $w \approx_\alpha v$ iff $K_\alpha\varphi \in w \Rightarrow \varphi \in v$ for every φ of $\mathcal{L}_I$.*
- *For $\alpha \in Ags$, R_α^I is a relation on W^{Λ_I} given by the rule: $wR_\alpha^I v$ iff $I_\alpha\varphi \in w \Rightarrow \varphi \in v$ for every φ of $\mathcal{L}_I$.*
- *Recall that P is the set of propositions in $\mathcal{L}_I$. Then $\mathcal{V} : P \to 2^{W^{\Lambda_I}}$ is the canonical valuation, defined so that $w \in \mathcal{V}(p)$ iff $p \in w$.*

Proposition D.23. *The canonical structure $\mathcal{M}$ for Λ_I is a Kripke-*ies*-model.*

Proof. We want to show that the tuple $\left\langle W^{\Lambda_I}, R_\Box, \texttt{Choice}, \{\approx_\alpha\}_{\alpha\in Ags}, \left\{\mathrm{R}^I_\alpha\right\}_{\alpha\in Ags}\right\rangle$ is a Kripke-*ies*-frame, which amounts to showing that the tuple satisfies the items in the definition of Kripke-*ies*-frames (Definition D.14).

1. The fact that $\left\langle W^{\Lambda_I}, R_\Box, \texttt{Choice}, \{\approx_\alpha\}_{\alpha\in Ags}\right\rangle$ meets the items of Definition C.41 is shown exactly as in Proposition C.48 (p. 207).

2. Since Λ_I includes the **KD45** schemata for I_α (for each $\alpha \in Ags$), then R^I_α is a serial, transitive, and euclidean relation on W for every $\alpha \in Ags$. Since Λ_I includes schema (*InN*), then $R^I_\alpha \subseteq\approx_\alpha \circ R_\Box$ for every $\alpha \in Ags$.

   We now verify that frame condition $(\texttt{Den})_\texttt{K}$ is satisfied. Take $\alpha \in Ags$. Let $v, u \in W^{\Lambda_I}$ be such that $v \approx_\alpha \circ R_\Box u$. This means that there exists $w \in W$ such that $v \in \overline{w}$ and $w \approx_\alpha u$. We want to show that there exists $z \in W$ such that $uR^I_\alpha z$ and $vR^I_\alpha z$. We show that $z' = \{\psi; I_\alpha\psi \in v\} \cup \{\psi; I_\alpha\psi \in u\}$ is consistent. To do so, we first show that $\{\psi; I_\alpha\psi \in v\}$ is consistent. Suppose for a contradiction that it is not consistent. Then there exists a set $\{\psi_1, \dots, \psi_n\}$ of formulas of $\mathcal{L}_\mathsf{I}$ such that $I_\alpha\psi_i \in v$ for every $1 \le i \le n$ and (a) $\vdash_{\Lambda_I} \psi_1 \wedge \dots \wedge \psi_n \to \bot$. By Necessitation for I_α and its distributivity over conjunction, (a) implies that $\vdash_{\Lambda_I} I_\alpha\psi_1 \wedge \dots \wedge I_\alpha\psi_n \to I_\alpha\bot$, but this is a contradiction, since v is a Λ_I-MCS and it includes $I_\alpha\psi_1 \wedge \dots \wedge I_\alpha\psi_n$. One can use an analogous argument to show that $\{\psi; I_\alpha\psi \in u\}$ is consistent. Next, we show that $z' = \{\psi; I_\alpha\psi \in v\} \cup \{\psi; I_\alpha\psi \in u\}$ is also consistent. Suppose for a contradiction that it is not consistent. Since $\{\psi; I_\alpha\psi \in v'\}$ and $\{\psi; I_\alpha\psi \in u\}$ are consistent, then there must exist sets $\{\varphi_1, \dots, \varphi_n\}$ and $\{\theta_1, \dots, \theta_m\}$ of formulas of $\mathcal{L}_\mathsf{I}$ such that $I_\alpha\varphi_i \in v$ for every $1 \le i \le n$, $I_\alpha\theta_i \in u$ for every $1 \le i \le m$, and (b) $\vdash_{\Lambda_I} (\varphi_1 \wedge \dots \wedge \varphi_n) \wedge (\theta_1 \wedge \dots \wedge \theta_m) \to \bot$. Let $\theta = \theta_1 \wedge \dots \wedge \theta_m$ and $\varphi = \varphi_1 \wedge \dots \wedge \varphi_n$. Thus, (b) implies that $\vdash_{\Lambda_I} \varphi \to \neg\theta$ and thus that (c) $\vdash_{\Lambda_I} \langle I_\alpha\rangle\varphi \to \langle I_\alpha\rangle\neg\theta$. Notice that the facts that $I_\alpha\varphi_i \in v$ for every $1 \le i \le n$, that I_α distributes over conjunction, and that v is a Λ_I-MCS imply that $I_\alpha\varphi \in v$. Analogously, one has that ($\star$) $I_\alpha\theta \in u$. The fact that $v \in \overline{w}$ implies that $\Diamond I_\alpha\varphi \in w$, so that (*Den*) entails that $K_\alpha\langle I_\alpha\rangle\varphi \in w$. Now, this last inclusion implies, with the fact that $w \approx_\alpha u$, that $\langle I_\alpha\rangle\varphi \in u$, which by (c) in turn yields that $\langle I_\alpha\rangle\neg\theta \in u$, contradicting ($\star$). Therefore, z' is consistent. Let z be the Λ_I-MCS that includes z', which exists in virtue of Lindenbaum's Lemma (Blackburn et al., 2002, Chapter 4, p. 199). By construction, $uR^I_\alpha z$ and $vR^I_\alpha z$. With this, we have shown that $\mathcal{M}$ satisfies $(\texttt{Den})_\texttt{K}$.

□

Lemma D.24 (Existence for non-intentional operators)**.** *Let $\mathcal{M}$ be the canonical Kripke-*ies*-model for Λ_I. For every $w \in W^{\Lambda_I}$ and every φ of $\mathcal{L}_I$, the following items hold:*

1. *$\Box\varphi \in w$ iff $\varphi \in v$ for every $v \in \overline{w}$.*

2. *$[\alpha]\varphi \in w$ iff $\varphi \in v$ for every $v \in \overline{w}$ such that $wR_\alpha^{\overline{w}}v$.*

3. *$K_\alpha\varphi \in w$ iff $\varphi \in v$ for every $v \in W^{\Lambda_I}$ such that $w \approx_\alpha v$.*

Proof. The proof is the same as the one included for Lemma A.16 (p. 121). □

Lemma D.25 (Truth Lemma)**.** *Let $\mathcal{M}$ be the canonical Kripke-*ies*-model for Λ_I. For all φ of $\mathcal{L}_I$ and $w \in W^{\Lambda_I}$, $\mathcal{M}, w \models \varphi$* iff *$\varphi \in w$.*

Proof. We proceed by induction on the complexity of φ. The cases of propositional letters and of Boolean connectives are standard. For the cases of $\Box, [\alpha]$, and K_α, both directions follow straightforwardly from Lemma D.24 (items 1, 2, and 3, respectively). As for I_α, we have the following arguments:

- ("I_α") ($\Rightarrow$) We work by contraposition. Suppose that $I_\alpha\varphi \notin w$. Take $x \in \pi_\alpha^\Box[w]$. The assumption that $\neg I_\alpha\varphi \in w$ implies, by schema (*KI*) and closure of w under *Modus Ponens*, that $\Box K_\alpha \neg I_\alpha\varphi \in w$. Since $x \in \pi_\alpha^\Box[w]$, this implies that $\neg I_\alpha\varphi \in x$. By an argument analogous to the one used in Proposition D.23 to show that the canonical model satisfies $(\texttt{Den})_{\texttt{K}}$, the set $\{\psi; I_\alpha\psi \in x\}$ is consistent. Next, observe that $\{\psi; I_\alpha\psi \in w\} \cup \{\neg\varphi\}$ is consistent. To see this, suppose that it is not consistent. Since $\{\psi; I_\alpha\psi \in w\}$ is consistent, there must exist a set $\{\varphi_1, \ldots, \varphi_n\}$ such that $I_\alpha\varphi_i \in w$ for every $1 \leq i \leq n$ and $\vdash_{\Lambda_I} (\varphi_1 \wedge \cdots \wedge \varphi_n) \wedge \neg\varphi \to \bot$ Now, this Λ_I-theorem implies that $\vdash_{\Lambda_I} (\varphi_1 \wedge \cdots \wedge \varphi_n) \to \varphi$. By Necessitation of I_α, its schema (*K*), and its distributivity over conjunction, one then has that ($\star$) $\vdash_{\Lambda_I} (I_\alpha\varphi_1 \wedge \cdots \wedge I_\alpha\varphi_n) \to I_\alpha\varphi$. Now, closure of w under conjunction then implies that $(\bigwedge_{1 \leq i \leq n} I_\alpha\varphi_i) \in x$, so that the antecedent in Λ_I-theorem ($\star$) lies in x. Closure of x under *Modus Ponens* then implies that $I_\alpha\varphi \in x$, but this contradicts the previously shown fact that $\neg I_\alpha\varphi \in x$. Therefore, $\{\psi; I_\alpha\psi \in w\} \cup \{\neg\varphi\}$ is in fact consistent. Let u be the Λ_I-MCS that includes $\{\psi; I_\alpha\psi \in w\} \cup \{\neg\varphi\}$, which exists in virtue of Lindenbaum's Lemma (Blackburn et al., 2002, Chapter 4, p. 199). By construction, $xR_\alpha^I u$, so that $u \in x \uparrow_{R_\alpha^+}$. It also follows from construction that $\neg\varphi \in x$, so that the induction hypothesis yields that $\mathcal{M}, x \models \neg\varphi$. Thus, x is such that $x \in \pi_\alpha^\Box[w]$ and such that $x \uparrow_{R_\alpha^+} \nsubseteq |\varphi|$, which implies that $\mathcal{M}, w \not\models I_\alpha\varphi$.

  ($\Leftarrow$) Assume that $I_\alpha\varphi \in w$. Suppose for a contradiction that $\mathcal{M}, w \not\models I_\alpha\varphi$. This means that for all $x \in \pi_\alpha^\Box[w]$ there exists $y \in W$ such that $xR_\alpha^{I+}y$ and $\mathcal{M}, y \not\models \varphi$. Now, we have two cases. Case 1 comes from assuming that for every $x \in \pi_\alpha^\Box[w]$ the y such that $xR_\alpha^{I+}y$ and $\mathcal{M}, y \not\models \varphi$ is actually x itself. In this case, $\mathcal{M}, x \not\models \varphi$ for every $x \in \pi_\alpha^\Box[w]$. By induction hypothesis, this implies that $\neg\varphi \in x$ for every $x \in \pi_\alpha^\Box[w]$, which, by items 1 and 3 of Lemma D.24, implies that $\Box K_\alpha \neg\varphi \in x$ for every $x \in \pi_\alpha^\Box[w]$. In particular, $\Box K_\alpha \neg\varphi \in w$. Schema (*InN*) and closure of w under *Modus Ponens* then imply that $I_\alpha \neg\varphi \in w$, but this is a contradiction, since the fact that $I_\alpha\varphi \in w$, with schema (*D*) for I_α and closure of w under *Modus Ponens*, implies that $\neg I_\alpha \neg\varphi \in w$. Case 2 comes from assuming that that there exist $x \in \pi_\alpha^\Box[w]$ and $y \in W$ such that $xR_\alpha^{I+}y$, $\mathcal{M}, y \not\models \varphi$, and $y \neq x$. By induction hypothesis, $\varphi \notin y$. Since $xR_\alpha^{I+}y$ and $y \neq x$, then $xR_\alpha^{I}y$, so that the definition of R_α^I implies that $I_\alpha\varphi \notin x$. As such, $\neg I_\alpha\varphi \in x$, which, by schema (*KI*) and closure of x under *Modus Ponens*, implies that $\Box K_\alpha \neg I_\alpha\varphi \in x$. Since $x \in \pi_\alpha^\Box[w]$, this implies that $\neg I_\alpha\varphi \in w$, but this is a contradiction to the initial assumption.

□

Proposition D.26 (Completeness w.r.t. Kripke-*ies*-models). *The proof system Λ_I is complete with respect to the class of Kripke-*ies*-models.*

Proof. Let φ be a Λ_I-consistent formula of $\mathcal{L}_\mathsf{I}$. Let w be the Λ_I-MCS including φ, which exists in virtue of Lindenbaum's Lemma (Blackburn et al., 2002, Chapter 4, p. 199). Then the canonical structure $\mathcal{M}$ for Λ_I is a Kripke-*ies*-model such that $\mathcal{M}, w \models \varphi$, according to Lemma D.25 above. □

Proposition D.27 (Completeness w.r.t. *iebt*-models). *The proof system Λ_I is complete with respect to the class of* iebt-*models.*

Proof. Let φ be a Λ_I-consistent formula of $\mathcal{L}_\mathsf{I}$. Proposition D.26 implies that there exists a Kripke-*ies*-model $\mathcal{M}$ and a world w in its domain such that $\mathcal{M}, w \models \varphi$. Proposition D.21 then ensures that the *iebt*-model $\mathcal{M}^T$ associated with $\mathcal{M}$ is such that $\mathcal{M}^T, \langle \overline{w}, h_w \rangle \models \varphi$. □

Therefore, Proposition D.13 and Proposition D.27 imply that the following result, appearing in the main body of the chapter, has been shown:

Theorem 5.10 (Soundness & Completeness of Λ_I). *The proof system Λ_I is sound and complete with respect to the class of* iebt-*models.*

□

6

Responsibility

But since moral activity is unthinkable without physical activity, the cause of an event is neither the one nor the other, but a combination of the two.

Leo Tolstoi, *War and Peace*

'There is only one way to salvation, and that is to make yourself responsible for all men's sins. As soon as you make yourself responsible in all sincerity for everything and for everyone, you will see at once that this is really so, and that you are in fact to blame for everyone and for all things.'

'But then there are the children... if it is really true that they must share responsibility for all their fathers' crimes, such a truth is not of this world and is beyond my comprehension.'

Fyodor Dostoevsky, *Brothers Karamazov*

The study of responsibility is a complicated matter. The term is used in different ways in different fields, and it is easy to engage in everyday discussions as to why someone should be considered responsible for something. Typically, the backdrop of these discussions involves social, legal, moral, or philosophical problems, each with slightly different meanings for expressions like *being responsible for...*, *being held responsible for...*, or *having the responsibility of...*, among others. Therefore—to approach such problems efficiently—there is a demand for clear, taxonomical definitions of responsibility.

For instance, suppose that you are a judge in Texas. You are presiding over a trial where the defendant is being charged with first-degree murder. The alleged crime is horrible, and the prosecution seeks capital punishment. The case is as follows: driving her car, the defendant ran over a traffic officer that was holding a stop-sign at a crossing walk, while school children were crossing the street. The traffic officer was killed, and some of the children were severely injured. A highly complicated case, the possibility of a death-penalty sentence means that the life of the defendant is at stake. More than ever, due process is imperative. As the presiding judge, you must abide by the prevailing definitions of criminal liability with precision. In other words, there is little to no room for ambiguity in the ruling, and your handling of the notions associated with responsibility in criminal law should be impeccable.

As this example suggests, a framework with intelligible, realistically applicable definitions of responsibility is paramount in the field of law. However, responsibility-related problems arise across many other disciplines—social psychology, philosophy of emotion, legal theory, and ethics, to name a few (Lorini et al., 2014; Weiner, 1995). A clear pattern in all these is the intent of issuing standards for when—and to what extent—an agent should be held responsible for a state of affairs.

This is where Logic lends a hand. The development of expressive logics—to reason about agents' decisions in situations with moral consequences—involves devising unequivocal representations of components of behavior that are highly relevant to systematic responsibility attribution and to systematic blame-or-praise assignment. To put it plainly, expressive syntactic-and-semantic frameworks help us analyze responsibility-related problems in a methodical way.[1] Most likely, this is why the logic-based formalization of responsibility has become such an important topic in, for instance, normative multi-agent systems, responsible autonomous agents, and machine ethics for AI (Arrieta et al., 2020; Pereira & Saptawijaya, 2016).

In Chapter 1 I stated that the main goal of the whole thesis is to develop a formal theory of responsibility. Chapters 2–5 were key steps toward this aim, and here I finally present my proposal. As also mentioned in Chapter 1, this proposal

[1] Indeed, to produce the aforementioned 'unequivocal representations' is one of the most fundamental premises of applying game theory, decision theory, and deontic logic in the task of systematizing responsibility attribution. The previous chapters include several examples of multi-agent decision contexts, where the effects of choices of agents have ethical implications. It is precisely because of these ethical implications that the examples illustrate some of the typical problems that researchers want to address when it comes to producing a theory of responsibility.

relies on (a) a *decomposition* of responsibility into specific components and (b) a functional *classification* of responsibility, where the different categories directly correlate with the components of the decomposition.

As for the decomposition, the components of responsibility will at this point come as no surprise to the reader, since each accounts for a major topic of study in this thesis:

- **Agency**: the process by which agents bring about states of affairs in the environment. In other words, the phenomenon by which agents choose and perform actions, with accompanying mental states, that change the environment.
- **Knowledge and belief**: mental states that concern the information available in the environment and that explain agents' particular choices of action.
- **Intentions**: mental states that determine whether an action was done with the purpose of bringing about its effects.
- **Ought-to-do's**: the actions that agents should perform, complying to the codes of a normative system. Oughts-to-do's make up contexts that provide a criterion for deciding whether an agent should be blamed or praised. I refer to these contexts as the *deontic contexts* of responsibility.

As for the classification, it is a refinement of Broersen's three categories of responsibility: *causal*, *informational*, and *motivational* responsibility (see, for instance, Ågotnes, 2006;Broersen, 2008a; Duijf, 2018, Introduction). I will discuss these categories at length in Section 6.1.

On the basis of both the decomposition and the classification, in this chapter I introduce a very rich stit logic to analyze responsibility, which I refer to as *intentional epistemic act-utilitarian stit theory* (*IEAUST*). More precisely, I use *IEAUST* to model and syntactically characterize various modes of responsibility. By 'modes of responsibility' I mean combinations of sub-categories of the three ones mentioned above, cast against the background of particular deontic contexts.

Let me clarify. On the one hand, the sub-categories correspond to the different versions of responsibility that one can consider according to the *active* and *passive* forms of the notion: while the active form involves contributions—in terms of explicitly bringing about outcomes—the passive form involves omissions—which are interpreted as the processes by which agents allow that an outcome happens while being able, to some extent, to prevent it. On the other hand, the deontic context of a mode establishes whether and to what degree the combination of sub-categories involves either blameworthiness or praiseworthiness.

Now, the logic *IEAUST* is obtained by merging all the frameworks of the previous chapters. Thus, the formalism includes a language that expresses agency, epistemic notions, intentionality, and different senses of obligation. With this language, I characterize the components of responsibility using particular formulas. Then, adopting a compositional approach—where complex modalities are built out of more basic ones—I use these characterizations of the components to formalize the aforementioned modes of responsibility. An outline of this chapter is included below.

- Section 6.1 presents an operational definition for responsibility and addresses the philosophical perspective adopted in my study of the notion. The section discusses (a) backward- and forward-looking responsibility, (b) Broersen's three categories of responsibility, and (c) the active and passive forms for particular instances of these three categories. This conceptual analysis serves as groundwork for the stit-theoretic formalizations that the chapter investigates later on.

- Section 6.2 discusses *blame* and *praise* as central elements of my operational definition of responsibility, relating these two concepts to the deontic modalities of Chapter 4.

- Section 6.3 introduces *IEAUST* and uses this logic to provide stit-theoretic characterizations of different modes of responsibility (according to Broersen's categories of responsibility in their active and passive forms). These characterizations are illustrated with typical stit-like examples.

- Section 6.4 briefly reviews important logic-based properties of *IEAUST*. As for metalogic properties, the section presents Hilbert-style proof systems both for *IEAUST* and for a technical extension, addressing the status of their soundness & completeness results.

- Section 6.5 (the conclusion) first examines an extension of *IEAUST* with a probabilistic semantics of belief and with doxastic obligations; then, it offers a proposal for characterizing the modes of *mens rea*; lastly, it mentions some possibilities for future work in the context of collective responsibility.

6.1 Categories of Responsibility

To make a start on formally analyzing responsibility, I identify (a) two *viewpoints* for the philosophical study of responsibility, (b) three main *categories* for the viewpoint that I focus on, and (c) two *forms* in which the elements of the categories can be interpreted. Therefore, I am merging ideas of Lorini (2013) (whose main categorization of responsibility comes from Aristotle's *Nicomachean Ethics*), of Broersen (2011a), of van de Poel (2011), and of Talbert (2019). Let me further explain points (a), (b), and (c).

As for (a), the philosophical literature on responsibility usually distinguishes two *viewpoints* on the notion (see van de Poel, 2011, for details): *backward-looking responsibility* and *forward-looking responsibility*. By backward-looking responsibility one refers to the viewpoint according to which an agent is considered to have produced a state of affairs that has already ensued and lies in the past. This is the viewpoint taken by a judge when, while trying a murder case, she wants to get to the bottom of things and find out who is responsible for doing the killing. In contrast, by forward-looking responsibility one refers to the viewpoint according to which which an agent is expected to comply with the duty of bringing about a state of affairs in the future. When one thinks of a student that has to write an essay before its due date, for instance, this is the view that is being used. In other words, the writing and the handing in of the essay before the deadline are seen as responsibilities of the student.

From here on, I will focus on backward-looking responsibility. Thus, unless explicitly stated otherwise, whenever the word 'responsibility' is used, I am referring to backward-looking responsibility. Even with this restriction, responsibility is immensely multifaceted. Just as stated in Chapter 1, then, I work with the following operational definition:

> ***Responsibility:*** *a relation between the agents and the states of affairs of an environment, such that an agent is responsible for a state of affairs iff the agent's degree of involvement in the realization of that state of affairs warrants blame or praise (in light of a given normative system).*[2]

As for point (b), I follow Broersen (2011a) (see also Duijf, 2018) and distinguish three main *categories* of responsibility, where each category can be correlated with the components of responsibility that it involves:

1. ***Causal responsibility***: an agent is causally responsible for a state of affairs iff the agent is the material author of such a state of affairs. The component that this category involves is agency.

[2] As stated by (Watson, 2001) and by (J. M. Fischer, 1982), for instance, the typical frameworks for analyzing responsibility deal with agents' responsibility for their *actions*, whereas agents can in principle also be seen as bearing responsibility for *omissions*, for *consequences*, and even for *character*. Making room for these distinctions, the logic that I introduce here accommodates not only agents' responsibility for actions but also responsibility for omissions and responsibility for consequences.

2. ***Informational responsibility***: an agent is informationally responsible for a state of affairs iff the agent is the material author and it behaved knowingly, or consciously, while bringing about the state of affairs. The components that this category involves are agency, knowledge, and belief.

3. ***Motivational responsibility***: an agent is motivationally responsible for a state of affairs iff the agent is the material author and it behaved knowingly and intentionally while bringing about the state of affairs. The components that this category involves are agency, knowledge, and intentions.

These categories extend the literature's common distinction between *causal* and *agentive* responsibility (see, for instance, Crisp, 2014; Lorini et al., 2014; Watson, 1996), and they were derived by Broersen on the basis of his analysis of the modes of *mens rea*.[3] Let me briefly elaborate on this matter. In American criminal law, the mental states that accompany an instance of *actus reus* (Latin for 'guilty act') are known as *mens rea* (Latin for 'guilty mind'). There are different *mens rea* mental states, and—as taken from(Dubber, 2002, pp. 62–80)—they correspond to the following levels of culpability, presented in decreasing order of culpability:[4]

- *Purposefully*: the actor has the 'conscious object' of engaging in conduct and believes and hopes that the attendant circumstances exist.

- *Knowingly*: the actor is certain that his conduct will lead to the result.

- *Recklessly*: the actor is aware that the attendant circumstances exist, but nevertheless engages in the conduct that a 'law-abiding person' would have refrained from.

- *Negligently*: the actor is unaware of the attendant circumstances and the consequences of his conduct, but a 'reasonable person' would have been aware.

- *Strict liability*: the actor engaged in conduct and his mental state is irrelevant.

[3] In fact, the modes of responsibility that are formalized in Section 6.3 are deeply connected with the modes of *mens rea* underlying Broersen's proposal. This means that, although my framework does not pretend to be applied in legal theory, it is certainly inspired by it (see Subsection 6.5.2 of this chapter's conclusion).

[4] It might be misleading to say 'decreasing order of culpability,' so let me clarify what this means. As presented by Broersen (2011a), it is not that a guilty act that is found out to be done *purposefully* yields a higher degree of culpability for an actor than a guilty act that is charged under *criminal negligence*. Rather, when considering a single guilty act, this is deemed to yield a higher degree of culpability if it was done purposefully rather than if the same act was done only with negligence. I will elaborate on this topic in Section 6.3 (when I formalize the modes of responsibility) and in this chapter's conclusion (Subsection 6.5.2).

As established in Chapter 1, Broersen realized that categorizing responsibility helps in the systematic identification of the differences between the modes of *mens rea*. This is the main idea behind his proposal of the three categories and behind the decomposition in this chapter's introduction (p. 263, see also the list of components on p. 3).

Finally, as for point (c), the two *forms* of responsibility are the *active* form and the *passive* form. The active form of causal responsibility concerns causal contributions: an agent is causal-active responsible for φ only if it saw to it that φ. The passive form of causal responsibility concerns causal omissions: an agent is causal-passive responsible for φ only if φ was the case, the agent could have prevented φ, but the agent refrained from preventing φ. The active form of informational responsibility concerns conscious contributions: an agent is informational-active responsible for φ only if it knowingly saw to it that φ. Its passive form concerns conscious omissions: an agent is informational-passive responsible for φ only if φ was the case, the agent knew that it could have prevented φ, and the agent knowingly refrained from preventing φ. The active form of motivational responsibility concerns motivational contributions: an agent is motivational-active responsible for φ only if it knowingly and intentionally brought about φ. Its passive form concerns motivational omissions: an agent is motivational-passive responsible for φ only if φ was the case and the agent knowingly and intentionally refrained from preventing φ.

This concludes my discussion on the philosophical analysis of responsibility and its categories. But why is one interested in these categories, in the first place? Before diving into formalizations, let me first open a new section to give an answer to this question.

6.2 Blame & Praise

Key elements in p. 265's operational definition of responsibility are the notions of blame and praise. Intuitively, responsibility can be measured by how much blame or how much praise an agent gets for its participation in bringing about a state of affairs. In other words, blame and praise are indicators of how responsible an agent should be held. A murderer, for instance, generally gets more blame—for the committed murder—than any of her accomplices. Similarly, when within a champion-team in the NBA an individual player's performance is ranked as most important for the team's success, that player gets the extra accolade of 'most valuable player.' These two examples illustrate that it is natural to think that degrees of praiseworthiness/blameworthiness are correlated with degrees of responsibility.

Nevertheless, blame and praise are more than indicators of responsibility. They are the main reasons for wanting to find out whether—and to what extent—an agent is responsible for a state of affairs. Let me explain this point. It is clear that there are actions that individuals/societies deem either reprehensible or commendable. Typically, these individuals/societies have an interest in figuring out who is responsible—and, if so, how much—for such actions, precisely so that the authors can be either sanctioned or honored. Not every action elicits this interest, however. While a mass murder almost demands that society finds out who was responsible for it, there would likely be no interest in finding out who made a particular footprint on a road that is busy on a daily basis. The discrepancy is due to the general condemnation of the act of murder, coupled with the prevalence of schemes of accountability for blameworthy agents in most legal societies. At the other end of the spectrum, finding out who is responsible for actions that call for praise is also standard. Consider the act of developing the cure for a deadly disease, for instance; to decorate whoever is responsible is part of a scheme in which prizes are awarded to authors, precisely to encourage acts that are desirable for society.[5]

Up to now, I have not explicitly discussed *blame* and *praise*, so it seems fitting to do so. Intuitively, these terms refer to attitudes that agents have, toward themselves and others, that regard the undesirability, resp. desirability, of a state of affairs that was obtained as a result of some action. As the examples in the previous paragraph show, this undesirability, resp. desirability, can in principle be related to many other measures, such as individual/social preferences, individual/social utility, individual/social morality, individual/social commitments, and individual/social codes of conduct (or norms), to name a few. The basic idea, nonetheless, is that undesirable outcomes are reasons for blaming the authors, and desirable outcomes are reasons for praising them.[6]

In this thesis there are specific contexts that provide a criterion for deciding when agents should be blamed and when agents should be praised. These contexts are given by the deontic attributes of the logics presented in Chapter 4. More precisely, consider the distinctions between an agent's available choices of

[5] That an agent is deserving of blame or praise is part of a complex feature typically attributed to human individuals and human groups known as the *morality of fairness* (see Tomasello, 2016, Chapter 1).

[6] It is in this sense that blame and praise can be seen as contrasting attitudes within the same spectrum. Whether blame and praise are magnitudes that are the exact opposite of each other is a further question that I do not attempt to answer here, however.

action, that are implied by the ought-to-do's that have been reviewed and formalized so far. Such distinctions, which in essence divide choices between optimal and non-optimal, can be used to build a framework that accounts for degrees of praiseworthiness/blameworthiness.

The main idea is as follows: if agent α ought to have done φ, then having seen to it that φ makes α praiseworthy, while having refrained from seeing to it that φ makes α blameworthy. For a given φ, then, the degrees of α's praiseworthiness/blameworthiness correspond to the possible combinations between (a) the deontic modalities introduced in Chapter 4 ($\odot_\alpha \varphi$ for objective ought-to-do's and $\odot_\alpha^S \varphi$ for subjective ought-to-do's), and (b) the active/passive forms of the three categories of responsibility. Although Section 6.3 discusses this in detail, let me clarify the intuition underlying such degrees of praiseworthiness/blameworthiness with some examples.

Suppose that you are driving through a quiet neighborhood, where there is always light traffic, in the middle of the day. You see that a few blocks down the road a traffic officer is holding up a stop-sign at a crossing walk so that a group of school children can cross the road behind her. Let s stand for the proposition 'your car is stopped.' In a situation like this, at all moments at which the school children are crossing, you objectively and subjectively ought to see to it that your car is stopped. Without going into details regarding epistemic indistinguishability, you know that the children are crossing the street—because you can see them. In terms of formulas and *bt*-models, formulas $\odot_{you} s$ and $\odot_{you}^S s$ hold at every index based on said moments. Suppose, nonetheless, that you knowingly and intentionally keep on driving. Thus, a formula of the form $K_{you}[you]\neg s \wedge I_{you}[you]\neg s$ holds. In this case, it would be difficult to say that you are not highly blameworthy for whatever terrible circumstance that ensues from your action.

In contrast, suppose that there is a terrorist inside your car, and that he is holding your family at gun point. This terrorist has threatened to shoot your family if you do not keep on driving. Making a tough choice, you keep on driving. Once again, *at all moments at which the kids are crossing* you objectively and subjectively ought to have stopped the car. However, although you knowingly kept on driving, you did not have any intention of doing so—and thus you did not intentionally keep on driving. Therefore, in this case formula $K_{you}[you]\neg s \wedge \neg I_{you} \neg s$ holds, and most people would say that your degree of blameworthiness is less than the one in the first scenario.

Finally, consider this alternate backstory. Suppose that, the night before, a terrorist hacked the wiring and the computer of your car. He also installed a camera on the front hood so that he would be able to see where you would be driving the next day. In the morning you noticed nothing strange about your car, and you started your drive as any other day. Right before you reached the crossing walk, the terrorist—who knew where you were because of the camera on the front hood—remotely changed the settings of your car so that the brake pedal turned into the accelerator and vice versa. At the crossing walk, then, you stepped on the brake pedal to stop the car, but to your dismay the car sped up instead, causing a tragedy. In this case, although you causally saw to it that the car still went ahead, it is clear that you did so unknowingly. Moreover, you did not have the intention of driving past the crossing walk when the kids were passing by. Thus, in this case formula $[you]\neg s \wedge \neg K_{you}[you]\neg s \wedge \neg I_{you}\neg s$ holds, and your degree of blameworthiness should be even less than the one in the second scenario.

Compare these driving scenarios with the *Miners Paradox*—the example opening Chapter 4 (p. 139). As mentioned in Example 4.10 (p. 153) and further explained in Chapter 4's Subsection 4.4.2, the rescuers objectively ought to have blocked a shaft of the mine: if b stands for the proposition 'a shaft is blocked,' then formula $\odot_{Res} b$ holds. However, blocking a shaft was not a subjective ought-to-do: formula $\neg \odot^{S}_{Res}\, b$ holds. Suppose, then, that the rescuers refrained from blocking a shaft, so that both mines were flooded and one miner drowned. Since the rescuers were not subjectively obligated to block a shaft, one can say that the degree of blameworthiness should be less than the one in a hypothetical case where blocking a shaft were also a subjective ought-to-do.

As the reader can see, these examples were focused only on blameworthiness. However, praiseworthiness can also be accounted for using the deontic modalities and the categories of responsibility. This is done by considering as a basic principle that agents that comply with their ought-to-do's are praiseworthy. Just as in the case of blameworthy actors, different combinations of modalities provide a background to reason about degrees of praiseworthiness, albeit from a slightly different perspective. To clarify, there are situations where the compliance with obligations rarely elicits high degrees of praise from society. Consider the example of stopping your car at the crossing walk to let a group of school children pass by. You objectively and subjectively ought to stop the car, and while the degree of blameworthiness is very high if you refrain from stopping the car, you would not receive a medal if you did stop it.

Still, for a given obligation—without specifying, for now, whether it is objective or subjective—the degree of praiseworthiness attributed to knowingly and intentionally complying with it, for instance, is intuitively higher than the one attributed to only knowingly doing so. The reason, just as in the crossing-walk example given for blameworthiness, is that somebody might be forcing you into doing something that you did not intend to do. For instance, suppose that you are the captain of a basketball team in the NBA. You have made it to the championship finals, and the series' last game is being played. There are only a few seconds on the clock, and your team has the final ball-possession. During a time-out, your team's coaches are deliberating what to do. As it happens, the head coach had previously designed a strategy for end-of-game plays, and this strategy involves your passing the ball to the strongest player in the team so that he takes the last shot. However, you are a bit of a 'ball hog,' and you would rather take all the credit by scoring yourself. During the time-out, then, you start an argument with the head coach, claiming that it is you who should take the shot. The coach threatens to fire you from the team if you do not follow his instructions, and all your teammates witness the argument. The play begins, and the ball gets to your hands. Very reluctantly, you pass the ball—knowingly complying with your obligation. Your team ends up winning the game, and you become an NBA champion. However, all your teammates know that you did not have any intention to pass the ball, and that you only did it because otherwise you would have been fired. Because of this, they do not praise your behavior whatsoever.[7]

In turn, the degree of praiseworthiness attributed to knowingly complying with an obligation is intuitively higher than the one attributed to complying with an obligation without knowing that one is doing such a thing. Another example from sports well illustrates this intuition. Suppose that you and a friend of yours are watching a dart-throwing tournament, where highly skilled players compete. As part of the entertainment, at some point the organizers will ask a person from the audience to come up and throw a single dart, blindfolded, so that if the person hits the bulls-eye they win a prize. Although to hit the bulls-eye can hardly be called an obligation, it is a 'deontically desirable' outcome—provided that no one wishes that the lucky contestant misses the shot. A random choice, the organizers ask the friend that you are with to be this lucky contestant. You know that he does not know the first thing about dart-throwing. Still, when he throws the dart, he hits the bulls-eye—a fluke. Without knowing how to throw a dart, and without

[7] Observe that, in this case, your ought-to-do concerns passing the ball instead of winning the game on your own. A reader that has carefully followed Chapter 4 might object and say that the action of passing the ball should have a lower utility than that of winning the game on your own. Therefore, in act-utilitarian stit theory, the action of passing the ball should not dominate that of taking the last shot—provided that there is a history within the latter choice where you take all the credit for winning the game by scoring yourself. According to this view, then, to pass the ball should not be your ought-to-do. In my view, the best response to this possible objection is that, in this example, *individual* utility does not match with *deontic* utility. Although your individual payoff will be greater if you score the last shot, this does not mean that the team's payoff will also be greater. The team's payoff is still the same: becoming champions. Therefore, if one considers deontic utility as the utility of the team instead of that of the individual players, one can argue that the action of passing the ball indeed dominates that of making the last shot.

even seeing the target, your friend managed. Therefore, although you might be pleasantly surprised, you would not praise your friend for his skill at throwing darts.

These examples suggest that a taxonomy where the possible conjunctions of modalities $\odot_\alpha\varphi$ and $\odot^S_\alpha\varphi$, on the one hand, with modalities $[\alpha]\varphi$, $K_\alpha[\alpha]\varphi$, and $I_\alpha\varphi$, on the other, can provide a criterion that accounts for degrees of praiseworthiness/blameworthiness. The next section includes an extensive discussion on such possibilities.

6.3 A Logic of Responsibility

We are ready to introduce *intentional epistemic act-utilitarian stit theory* (*IEAUST*), a stit-theoretic logic of responsibility. The reader will see that all the concepts addressed in previous chapters will greatly pay off at this point. The reason is that *IEAUST* is obtained by integrating (a) Chapter 4's epistemic act-utilitarian stit theory (*EAUST*) and (b) Chapter 5's intentional epistemic stit theory (*IEST*). The resulting logic might seem crammed, but the intuitions behind each modality have all been explained by now. Without further ado, let me address the syntax and semantics of this expressive framework.

6.3.1 Syntax & Semantics

Definition 6.1 (Syntax of intentional epistemic act-utilitarian stit theory). *Given a finite set Ags of agent names and a countable set of propositions P, the grammar for the formal language $\mathcal{L}_R$ is given by*

$$\varphi ::= p \mid \neg\varphi \mid \varphi \wedge \varphi \mid \Box\varphi \mid [\alpha]\varphi \mid K_\alpha\varphi \mid I_\alpha\varphi \mid \odot_\alpha\varphi \mid \odot^S_\alpha\varphi,$$

where p ranges over P and α ranges over Ags.

In this language, $\Box\varphi$ is meant to express the historical necessity of φ ($\Diamond\varphi$ abbreviates $\neg\Box\neg\varphi$); $[\alpha]\varphi$ expresses that 'agent α has seen to it that φ'; $K_\alpha\varphi$ expresses that 'α knows φ'; $I_\alpha\varphi$ expresses that 'α p-d intended φ'; $\odot_\alpha\varphi$ expresses that 'α objectively ought to have seen to it that φ'; and $\odot^S_\alpha\varphi$ expresses that 'α subjectively ought to have seen to it that φ.'[8]As for the semantics, the structures on which the

[8] Just as in all the other chapters of this thesis, the present description of the stit-theoretic modalities follows *my interpretation* of the semantics (see the discussion on p. 34 and Remark 2.4, p. 36). Therefore, when specifying the points of evaluation for the formulas—the indices in *bt*-models—I take it that at those indices states of affairs are definitive. Because of this, I use the present-perfect tense for the description of modality $[\alpha]\varphi$ and say that 'at index $\langle m, h\rangle$ α has seen to it that φ.' To be consistent, I

formulas of $\mathcal{L}_{\mathsf{R}}$ are evaluated are based on what I call *knowledge-intentions-oughts branching-time frames*. These are the usual *bt*-frames, supplemented with the functions and relations that make up the semantic counterpart of the modalities in $\mathcal{L}_{\mathsf{R}}$. Let me first present the formal definition of these frames and then review the intuitions behind the extensions.

Definition 6.2 (*Kiobt*-frames & models)**.** *A tuple*

$$\left\langle M, \sqsubset, Ags, \mathbf{Choice}, \{\sim_\alpha\}_{\alpha \in Ags}, \tau, \mathbf{Value} \right\rangle$$

is called a knowledge-intention-oughts branching-time frame *(*kiobt*-frame for short) iff*

- $\left\langle M, \sqsubset, Ags, \mathbf{Choice}, \{\sim_\alpha\}_{\alpha \in Ags}, \mathbf{Value} \right\rangle$ *is an* aubt*-frame (Definition 4.3).*
- $\left\langle M, \sqsubset, Ags, \mathbf{Choice}, \{\sim_\alpha\}_{\alpha \in Ags}, \tau \right\rangle$ *is an* iebt*-frame (Definition 5.4).*

A kiobt*-model $\mathcal{M}$, then, results from adding a valuation function $\mathcal{V}$ to a* kiobt*-frame, where $\mathcal{V} : P \to 2^{I(M \times H)}$ assigns to each atomic proposition a set of indices.*

Thus, a *kiobt*-frame is the extension of an *aubt*-frame with the function τ from Section 5.1, that assigns to each agent and index the topology of the agent's p-d intentions at that index. As is customary, these models are used to define semantics for the formulas of $\mathcal{L}_{\mathsf{R}}$. Before presenting these semantics, a brief review—both of concepts and of terminology—seems suitable. Let $\mathcal{M}$ be a *kiobt*-frame. Recall the following characteristics of its components:

- $\langle M, \sqsubset, Ags, \mathbf{Choice} \rangle$ is a *bt*-frame just as in Definition 2.2 (p. 29).
- For $\alpha \in Ags$, the equivalence relation $\sim_\alpha$ is the usual indistinguishability relation, borrowed from epistemic logic, that represents α's uncertainty in *ebt*-frames (Definition 2.27, p. 70). For each $\alpha \in Ags$, $\sim_\alpha$ satisfies conditions $(\mathtt{OAC})$ *own action condition* and $(\mathtt{Unif-H})$ *uniformity of historical possibility* (see Definition 4.18, p. 162). For $\alpha \in Ags$ and index $\langle m, h \rangle$, $\pi^{\Box}_{\alpha}[\langle m, h \rangle] = \{\langle m', h' \rangle; \exists h'' \in H_{m'} s.t. \langle m, h \rangle \sim_\alpha \langle m', h'' \rangle\}$ is known as α's ex ante *information set*; and $\pi_\alpha[\langle m, h \rangle] = \{\langle m', h' \rangle; \langle m, h \rangle \sim_\alpha \langle m', h' \rangle\}$ is known as α's ex interim *information set*.

use the past tense for modalities $\Box\varphi$, $K_\alpha\varphi$, and $I_\alpha\varphi$ and say that 'at index $\langle m, h \rangle$ φ was settled,' that 'at index $\langle m, h \rangle$ α knew φ,' and that 'at index $\langle m, h \rangle$ α p-d intended φ.' For modalities involving the verb 'ought,' I use the past form of the sentences and say that 'at index $\langle m, h \rangle$ α ought to have seen to it that φ.' For the same reason, I will use the past tense when describing α's responsibility for φ and say that 'at index $\langle m, h \rangle$ α was responsible for φ' (see p. 277). As discussed in Chapter 2, this usage does not mean to refer to past moments. Rather, it aims to reinforce the notion that, at the level of indices, circumstances in the world are definitive, have already happened, and cannot be changed.

- τ is a function that assigns to each $\alpha \in Ags$ and index $\langle m, h\rangle$ the topology $\tau_{\alpha}^{\langle m,h\rangle}$ of α's intentionality at $\langle m, h\rangle$, where any open set is interpreted as a p-d intention of α at $\langle m, h\rangle$. Function τ satisfies conditions (`CI`) *consistency of intention* and (`KI`) *knowledge of intention* (see Definition 5.4, p. 226).

- **Value** is a function that assigns to each history h a real number, representing the deontic utility of h. This function allows us to define objective and subjective orderings on an agent's choices of action. Recall from Definition 4.4 (p. 148) that, for $m \in M$ and $\beta \in Ags$, $\mathbf{State}_{\beta}^{m} := \left\{S \subseteq H_m; S = \bigcap_{\alpha \in Ags - \{\beta\}} s(\alpha), \text{ for } s \in \mathbf{Select}^m\right\}$, where $\mathbf{Select}^m$ denotes the set of all selection functions at m. First, a general ordering $\leq$ is defined on 2^{H_m} by the rule: $X \leq Y$ iff $\mathbf{Value}(h) \leq \mathbf{Value}(h')$ for every $h \in X$ and $h' \in Y$.

  For $\alpha \in Ags$ and $m \in M$, an objective dominance ordering $\preceq$ is then defined on $\mathbf{Choice}_{\alpha}^{m}$ by the rule: $L \preceq L'$ iff for all $S \in \mathbf{State}_{\alpha}^{m_*}, L \cap S \leq L' \cap S$. The objectively optimal set of actions is defined by $\mathbf{Optimal}_{\alpha}^{m} := \{L \in \mathbf{Choice}_{\alpha}^{m}; \text{there is no } L' \in \mathbf{Choice}_{\alpha}^{m} \text{ s. t. } L \prec L'\}$, where I write $L \prec L'$ iff $L \preceq L'$ and $L' \npreceq L$.

  Subjective orderings are defined using the notion of *epistemic clusters* (Definition 4.19, p. 163). For $\alpha \in Ags$, $m, m' \in M$, and $L \in \mathbf{Choice}_{\alpha}^{m}$, L's epistemic cluster at m' is the set $[L]_{\alpha}^{m'} := \{h' \in H_{m'}; \text{ there is } h \in L \text{ s. t. } \langle m, h\rangle \sim_{\alpha} \langle m', h'\rangle\}$. For $\alpha \in Ags$ and $m \in M$, an subjective dominance ordering $\preceq_s$ is then defined on $\mathbf{Choice}_{\alpha}^{m}$ by the rule: $L \preceq_s L'$ iff for all m' such that $m \sim_{\alpha} m'$ and $S \in \mathbf{State}_{\alpha}^{m'}$, $[L]_{\alpha}^{m'} \cap S \leq [L']_{\alpha}^{m'} \cap S$.[9] Just as in the case of objective ought-to-do's, this ordering allows us to define a subjectively optimal set of actions $\mathbf{SOptimal}_{\alpha}^{m} := \{L \in \mathbf{Choice}_{\alpha}^{m}; \text{ there is no } L' \in \mathbf{Choice}_{\alpha}^{m} \text{ s. t. } L \prec_s L'\}$, where I write $L \prec_s L'$ iff $L \preceq_s L'$ and $L' \npreceq_s L$.

Therefore, *kiobt*-frames allow us to represent the components of responsibility discussed in the introduction: agency, knowledge, intentions, and ought-to-do's. More precisely, they allow us to provide semantics for the modalities of $\mathcal{L}_{\mathsf{R}}$:

Definition 6.3 (Evaluation rules for *IEAUST*)**.** *Let $\mathcal{M}$ be a finite-choice* kiobt-*model, where I focus on finite-choice models to simplify the evaluation rules for objective and subjective ought-to-do's.*[10] *The semantics on $\mathcal{M}$ for the formulas of $\mathcal{L}_{\mathsf{R}}$ are recursively*

[9] Recall that I write $m \sim_{\alpha} m'$ if there exist $h \in H_m$ and $h' \in H_{m'}$ such that $\langle m, h\rangle \sim_{\alpha} \langle m', h'\rangle$.

[10] Finite-choice *bt*-models are those for which function **Choice** is such that $\mathbf{Choice}_{\alpha}^{m}$ is finite for every $\alpha \in Ags$ and $m \in M$ (see Footnote 17, p. 158). The reader is referred to Definitions 4.6 (p. 150) and 4.21 (p. 164) for the evaluation rules in the case of infinite-choice models.

defined by the following truth conditions:

$$\begin{array}{lll}
\mathcal{M}, \langle m,h\rangle \models p & \textit{iff} & \langle m,h\rangle \in \mathcal{V}(p) \\
\mathcal{M}, \langle m,h\rangle \models \neg\varphi & \textit{iff} & \mathcal{M}, \langle m,h\rangle \not\models \varphi \\
\mathcal{M}, \langle m,h\rangle \models \varphi \wedge \psi & \textit{iff} & \mathcal{M}, \langle m,h\rangle \models \varphi \textit{ and } \mathcal{M}, \langle m,h\rangle \models \psi \\
\mathcal{M}, \langle m,h\rangle \models \Box\varphi & \textit{iff} & \textit{for all } h' \in H_m, \mathcal{M}, \langle m,h'\rangle \models \varphi \\
\mathcal{M}, \langle m,h\rangle \models [\alpha]\varphi & \textit{iff} & \textit{for all } h' \in \mathbf{Choice}_\alpha^m(h), \mathcal{M}, \langle m,h'\rangle \models \varphi \\
\mathcal{M}, \langle m,h\rangle \models K_\alpha\varphi & \textit{iff} & \textit{for all } \langle m',h'\rangle \textit{ s. t. } \langle m,h\rangle \sim_\alpha \langle m',h'\rangle, \\
 & & \mathcal{M}, \langle m',h'\rangle \models \varphi \\
\mathcal{M}, \langle m,h\rangle \models I_\alpha\varphi & \textit{iff} & \textit{there exists } U \in \tau_\alpha^{\langle m,h\rangle} \textit{ s. t. } U \subseteq \|\varphi\| \\
\mathcal{M}, \langle m,h\rangle \models \odot_\alpha\varphi & \textit{iff} & \textit{for all } L \in \mathbf{Optimal}_\alpha^m, \mathcal{M}, \langle m,h'\rangle \models \varphi \\
 & & \textit{for every } h' \in L \\
\mathcal{M}, \langle m,h\rangle \models \odot_\alpha^S\varphi & \textit{iff} & \textit{for all } L \in \mathbf{SOptimal}_\alpha^m, \mathcal{M}, \langle m',h'\rangle \models \varphi \\
 & & \textit{for every } m' \textit{ s. t. } m \sim_\alpha m' \textit{ and every } h' \in [L]_\alpha^{m''}.
\end{array}$$

where $\|\varphi\|$ *refers to the set* $\{\langle m,h\rangle \in I(M \times H); \mathcal{M}, \langle m,h\rangle \models \varphi\}$. *I write* $\left|\varphi\right|^m$ *to refer to the set* $\{h \in H_m; \mathcal{M}, \langle m,h\rangle \models \varphi\}$.

6.3.2 Formalization of Sub-Categories of Responsibility

The logic introduced in the previous subsection allows us to formalize different modes of responsibility by means of formulas of $\mathcal{L}_{\mathsf{R}}$. Before diving into the formulas, let me present an operational definition for the expression 'mode of responsibility,' so that the reader has more clarity as to what I mean when I use it.

On the one hand, recall from Section 6.1 that my analysis of responsibility distinguishes three categories of the notion—Broersen's three categories of responsibility: *causal, informational,* and *motivational.* On the other hand, recall—also from Section 6.1—that such an analysis also presupposes two *forms* of responsibility: the active form—concerning contributions—and the passive form—concerning omissions. A mode of responsibility, then, is defined as follows: for $\alpha \in Ags$, index $\langle m,h\rangle$, and φ of $\mathcal{L}_{\mathsf{R}}$, a *mode of α's responsibility with respect to φ at $\langle m,h\rangle$* is a tuple consisting of three constituents: (1) a set of categories, taken from Broersen's three categories of responsibility, that applies to the relation between α and φ at $\langle m,h\rangle$, (2) the forms of responsibility—active or passive—that apply to the categories in said set, and (3) a deontic context, determining whether the forms of the categories are either blameworthy, praiseworthy, or neutral.

As for constituents (1) and (2), observe that the active and passive forms of the three categories of responsibility lead to sub-categories of the notion. For clarity, first I will introduce the stit-theoretic characterizations of these sub-categories;

afterwards, in Subsection 6.3.3, these sub-categories will be discussed against the backdrop of the deontic contexts that will decide their degree of blameworthiness or praiseworthiness (constituent (3) in a given mode).

Now, a maxim usually endorsed in the philosophical literature on moral responsibility is the *principle of alternate possibilities*. According to this principle, "a person is morally responsible for what he has done only if he could have done otherwise" (Frankfurt, 2018, p. 829).[11] In other words, an agent is not morally responsible for having brought about φ if φ was inevitable. Thus, the principle of alternate possibilities is clearly related to the notion of *deliberative agency* (Horty, 1989; Horty & Belnap, 1995; von Kutschera, 1986). Recall from Chapter 2's Definition 2.6 (p. 43) that the idea behind the deliberative-stit modality $[\alpha]^d\psi := [\alpha]\psi \wedge \Diamond\neg\psi$ is that if a state of affairs was already settled then the state did not really depend on an agent's choice of action and should not be thought of as an effect of any such choice. For $[\alpha]^d\psi$ to hold at some index, then, there are two requirements: (1) that ψ is an effect of the choice that α has performed at the index, known as the *positive condition*; and (2) that $\neg\psi$ must have been possible at the index, known as the *negative condition*.

Following the example of Lorini et al. (2014), I adopt the intuitions behind deliberative agency and restrict my view on responsibility to situations where agents can be said to actually have had a hand in bringing about states of affairs. Therefore, each sub-category of α's responsibility with respect to φ at $\langle m, h\rangle$ will include a positive condition—concerning the realization of φ—and a negative condition—concerning the realization of $\neg\varphi$. For $\alpha \in Ags$ and φ of $\mathcal{L}_{\mathsf{R}}$, the main sub-categories of α's responsibility with respect to φ are displayed in Table 6.1.

Form *Category*	Active (contributions)	Passive (omissions)
Causal	$[\alpha]\varphi \wedge \Diamond[\alpha]\neg\varphi$	$\varphi \wedge \Diamond[\alpha]\neg\varphi$
Informational	$K_\alpha[\alpha]\varphi \wedge \Diamond K_\alpha[\alpha]\neg\varphi$	$\varphi \wedge K_\alpha\neg[\alpha]\neg\varphi \wedge$ $\Diamond K_\alpha[\alpha]\neg\varphi$
Motivational	$K_\alpha[\alpha]\varphi \wedge I_\alpha[\alpha]\varphi \wedge$ $\Diamond K_\alpha[\alpha]\neg\varphi$	$\varphi \wedge K_\alpha\neg[\alpha]\neg\varphi \wedge$ $I_\alpha\neg[\alpha]\neg\varphi \wedge \Diamond K_\alpha[\alpha]\neg\varphi$

Table 6.1: *Main sub-categories.*

[11] It is worth mentioning that Frankfurt (2018) argued against this principle.

Let me explain and discuss Table 6.1. Let $\mathcal{M}$ be a *kiobt*-model. For $\alpha \in Ags$ and index $\langle m, h \rangle$, the sub-categories of α's responsibility with respect to φ at $\langle m, h \rangle$ are defined as follows:[12]

- - α was *causal-active responsible* for φ at $\langle m, h \rangle$ iff at $\langle m, h \rangle$ α has seen to it that φ (the positive condition) and it was possible for α to prevent φ (the negative condition). As such, I refer to state of affairs φ as a causal contribution of α at $\langle m, h \rangle$.[13]
  - α was *causal-passive responsible* for φ at $\langle m, h \rangle$ iff at $\langle m, h \rangle$ φ was the case (the positive condition), and α refrained from preventing φ while it was possible for α to prevent φ (the negative conditions). To clarify, formula $\varphi \to \neg[\alpha]\neg\varphi$ is valid, so that if φ was the case then α refrained from preventing φ. I refer to $\neg\varphi$ as a causal omission of α at $\langle m, h \rangle$.
- - α was *informational-active responsible* for φ at $\langle m, h \rangle$ iff at $\langle m, h \rangle$ α has knowingly seen to it that φ (the positive condition) and it was possible for α to knowingly prevent φ (the negative condition). I refer to φ as a conscious contribution of α at $\langle m, h \rangle$.
  - α was *informational-passive responsible* for φ at $\langle m, h \rangle$ iff at $\langle m, h \rangle$ φ was the case (the positive condition), and α knowingly refrained from preventing φ while it was possible for α to knowingly prevent φ (the negative conditions). I refer to $\neg\varphi$ as a conscious omission of α at $\langle m, h \rangle$.

[12] Recall from Chapter 2's Subsection 2.2.3 (p. 38) that the following expressions are standard in this thesis:

- *α has refrained from seeing to it that φ at $\langle m, h \rangle$ iff $\neg[\alpha]\varphi$ holds at $\langle m, h \rangle$.*
- *α has prevented φ at $\langle m, h \rangle$ iff $[\alpha]\neg\varphi$ holds at $\langle m, h \rangle$.*
- *α has refrained from preventing φ at $\langle m, h \rangle$ iff $\neg[\alpha]\neg\varphi$ holds at $\langle m, h \rangle$.*

Bringing knowledge and intentions into the picture, I will abide by the following conventions:

- *α has knowingly seen to it that φ at $\langle m, h \rangle$ iff $K_\alpha[\alpha]\varphi$ holds at $\langle m, h \rangle$.*
- *α has knowingly refrained from seeing to it that φ at $\langle m, h \rangle$ iff $K_\alpha\neg[\alpha]\varphi$ holds at $\langle m, h \rangle$.*
- *α has knowingly refrained from preventing φ at $\langle m, h \rangle$ iff $K_\alpha\neg[\alpha]\neg\varphi$ holds at $\langle m, h \rangle$.*
- *α has intentionally seen to it that φ at $\langle m, h \rangle$ iff $[\alpha]\varphi \land I_\alpha[\alpha]\varphi$ holds at $\langle m, h \rangle$.*
- *α has intentionally refrained from seeing to it that φ at $\langle m, h \rangle$ iff $\neg[\alpha]\varphi \land I_\alpha\neg[\alpha]\varphi$ holds at $\langle m, h \rangle$.*
- *α has intentionally refrained from preventing φ at $\langle m, h \rangle$ iff $\neg[\alpha]\neg\varphi \land I_\alpha\neg[\alpha]\neg\varphi$ holds at $\langle m, h \rangle$.*

[13] Observe that, in this case, α is the sole author of φ. In other words, α did not merely contribute to the realization of φ but is the primary and only reason why φ was realized. This follows from the fact that if $[\alpha]\varphi \land \Diamond[\alpha]\neg\varphi$ holds at $\langle m, h \rangle$ then no other agent could have seen to it that φ at such an index.

- - α was *motivational-active responsible* for φ at $\langle m,h\rangle$ iff at $\langle m,h\rangle$ α has both knowingly and intentionally seen to it that φ (the positive conditions) and it was possible for α to knowingly prevent φ (the negative condition). I refer to φ as a motivational contribution of α at $\langle m,h\rangle$.
  - α was *motivational-passive responsible* for φ at $\langle m,h\rangle$ iff at $\langle m,h\rangle$ φ was the case (the positive condition), and α both knowingly and intentionally refrained from preventing φ while it was possible for α to knowingly prevent φ (the negative conditions). I refer to $\neg\varphi$ as a motivational omission of α at $\langle m,h\rangle$.

The main reason for setting the negative conditions as stated in Table 6.1 is that it greatly simplifies the relation between the active and the passive forms of responsibility. Namely, it implies that passive responsibility is a logical consequence of active responsibility (see Observation 6.4 2), something that in turn simplifies Subsection 6.3.3's presentation of modes of responsibility. That said, it is important to mention that these negative conditions lead to a policy that I call *leniency on blameworthy agents*. Let me elaborate on this topic.

Consider the definition of causal responsibility. In the present framework, agent α was causal-active responsible for φ at $\langle m,h\rangle$ *only if* $[\alpha]^d\varphi$ holds at $\langle m,h\rangle$.[14] Similarly, α was causal-passive responsible for φ at $\langle m,h\rangle$ *only if* $\neg[\alpha]^d\neg\varphi\wedge\Diamond[\alpha]^d\neg\varphi$ holds at $\langle m,h\rangle$.[15] This is why my proposal can be called 'lenient on blameworthy agents': for α to be causal-active responsible for φ at $\langle m,h\rangle$, both $[\alpha]\varphi$ (the positive condition) and $\Diamond[\alpha]\neg\varphi$ (the negative condition) must hold at $\langle m,h\rangle$. As such, that $[\alpha]^d\varphi$ holds at $\langle m,h\rangle$, for example, is not enough to guarantee that α will be causal-active responsible for φ at $\langle m,h\rangle$. To clarify, $[\alpha]\varphi\wedge\Diamond\neg\varphi$ does not logically imply $[\alpha]\varphi\wedge\Diamond[\alpha]\neg\varphi$ in the present framework. For one to be casual-active responsible, then, it is not enough that one has seen to it that φ while $\neg\varphi$ is possible; one must have seen to it that φ while being able to prevent φ. I consider such a policy 'lenient on blameworthy agents' because the requirements for causal-active responsibility can be set using weaker formulas—precisely like $[\alpha]\varphi\wedge\Diamond\neg\varphi$. The same argument applies to α's causal-passive responsibility: one can use weaker formulas to define it, such as, for instance, $\varphi\wedge\Diamond\neg\varphi$.

[14] This follows from the validity of schema (D) for $[\alpha]$, modal logic, and the fact that $\Diamond\neg\varphi\leftrightarrow\Diamond\neg[\alpha]\varphi$ is valid with respect to all *bt*-frames. Namely, since $\Diamond[\alpha]\neg\varphi\to\Diamond\neg[\alpha]\varphi$ is valid, then the validity of $\Diamond\neg\varphi\leftrightarrow\Diamond\neg[\alpha]\varphi$ implies that $\Diamond[\alpha]\neg\varphi\to\Diamond\neg\varphi$ is also valid.

[15] This follows from the following arguments: since $\varphi\to\neg[\alpha]\neg\varphi$ is valid (from the validity of schema (T) for $[\alpha]$), then α was causal-passive responsible for φ at $\langle m,h\rangle$ only if $\neg[\alpha]\neg\varphi\wedge\Diamond[\alpha]\neg\varphi$ holds at $\langle m,h\rangle$—that is, only if α has *deliberatively refrained*, or von-Wright-refrained, from seeing to it that $\neg\varphi$ (see the discussion on refraining on p. 39); in light of the definition given by $[\alpha]^d\psi := [\alpha]\psi\wedge\Diamond\neg\psi$, formula $\neg[\alpha]^d\psi\wedge\Diamond[\alpha]^d\psi$ translates into a formula that is logically equivalent to $\neg[\alpha]\psi\wedge\Diamond[\alpha]\psi$.

As for informational responsibility, the policy of leniency on blameworthy agents carries on. Here, the positive condition is given by $K_\alpha[\alpha]\varphi$, and the negative condition is given by $\Diamond K_\alpha[\alpha]\neg\varphi$. Thus, that $K_\alpha[\alpha]^d\varphi$ holds at $\langle m, h\rangle$ is not enough to ensure that α will be informational-active responsible for φ at $\langle m, h\rangle$. To clarify, $K_\alpha[\alpha]^d\varphi$ translates into $K_\alpha[\alpha]\varphi \wedge K_\alpha\Diamond\neg\varphi$, and this formula does not logically imply $K_\alpha[\alpha]\varphi \wedge \Diamond K_\alpha\neg[\alpha]\varphi$ in the present framework. For one to be informational-active responsible, it is not enough that one has knowingly seen to it that φ while knowing that φ is possible; one must have knowingly seen to it that φ while being able to knowingly prevent φ. Just as in causal responsibility, the requirements for informational-active responsibility could in principle be set using far weaker formulas—such as, for instance, $K_\alpha[\alpha]\varphi \wedge \Diamond K_\alpha\neg[\alpha]\varphi$, or the weaker $K_\alpha[\alpha]\varphi \wedge K_\alpha\Diamond\neg[\alpha]\varphi$, or the even weaker $K_\alpha[\alpha]\varphi \wedge \Diamond\neg\varphi$. Thus, once again we find a policy of leniency on blameworthy agents, where the same argument applies to α's informational-passive responsibility: one can use weaker formulas to define it—for instance, $\varphi \wedge K_\alpha\neg[\alpha]\neg\varphi \wedge K_\alpha\Diamond[\alpha]\neg\varphi$, or, weaker than this one, either $\varphi \wedge K_\alpha\neg[\alpha]\neg\varphi \wedge \Diamond[\alpha]\neg\varphi$ or $\varphi \wedge K_\alpha\neg[\alpha]\neg\varphi \wedge K_\alpha\Diamond\neg\varphi$, or, weaker than both these two, $\varphi \wedge K_\alpha\neg[\alpha]\neg\varphi \wedge \Diamond\neg\varphi$.

Although based on the ideas behind deliberative agency, my policy of leniency on blameworthy agents might be considered too lenient by some. To illustrate this, consider Chapter 4's example in Figure 4.8 (p. 168), where *Nikolai* is gambling. Suppose that the actual index is $\langle m_2, h_3\rangle$, at which *Nikolai* has chosen the action of forfeiting the bet (N_3). Observe, then, that $\mathcal{M}, \langle m_2, h_3\rangle \models \neg\Diamond K_{Nik}[Nik]w \wedge K_{Nik}[Nik]\neg w$: *at* $\langle m_2, h_3\rangle$ *it was impossible for* Nikolai *to knowingly win, and he has knowingly lost.* According to Table 6.1, *Nikolai* would not be informational-active responsible for losing, since it was impossible for him to knowingly win. However, observe that $\mathcal{M}, \langle m_2, h_3\rangle \models \Diamond K_{Nik}(\neg K_{Nik}[Nik]\neg w \wedge \langle K_{Nik}\rangle [Nik]w)$ as well: *at* $\langle m_2, h_3\rangle$ *it was possible for* Nikolai *to choose an action for which he knew both that he was not knowingly losing by choosing it and that he had the epistemic possibility of winning.* To clarify, if *Nikolai* would have chosen either to bet heads (N_1) or to bet tails (N_2), then formula $\neg K_{Nik}[Nik]\neg w \wedge \langle K_{Nik}\rangle [Nik]w$ would hold, and schema (5) for K_{Nik} implies that $K_{Nik}(\neg K_{Nik}[Nik]\neg w \wedge \langle K_{Nik}\rangle [Nik]w)$ would hold as well. Thus, some people might object to the claim that *Nikolai* was not informational-active responsible for losing, since he chose an action where he knowingly lost when it was possible for him to choose an action for which he knew that he would not knowingly lose and that there was the epistemic possibility of winning. For those who wish to conclude that *Nikolai* is informational-active responsible for losing, they would get the desired result by revising the negative conditions of informational responsibility.

As for motivational responsibility, observe that the negative conditions—in both the active and the passive forms—involve epistemic attitudes. Recall from Chapter 5's Section 5.1 that p-d intentions are part of an agent's *ex ante* information set. Moreover, α's p-d intending φ implies that α knows that it is impossible to still have a p-d intention of $\neg\varphi$ at the same time—formula $I_\alpha\varphi \to K_\alpha\Box\neg I_\alpha\neg\varphi$ is valid.[16] Nevertheless, it still seems reasonable to impose negative conditions for motivational responsibility, just as in the other categories. I consider that, for α to be motivational-active responsible for φ at $\langle m,h\rangle$, at the very least α should have known that at $\langle m,h\rangle$ it was possible to refrain from seeing to it that φ—that is, $K_\alpha\Diamond\neg[\alpha]\varphi$ should hold. In turn, for α to be motivational-passive responsible for φ at $\langle m,h\rangle$, α should have knowingly and intentionally refrained from preventing φ and, at the very least, α should have known that at $\langle m,h\rangle$ it was possible to prevent φ—that is, $K_\alpha\neg[\alpha]\neg\varphi \wedge I_\alpha\neg[\alpha]\neg\varphi \wedge K_\alpha\Diamond[\alpha]\neg\varphi$ should hold. The policy of leniency on blameworthy agents, then, led me to set the requirements for motivational responsibility using negative conditions similar to those in the category of informational responsibility: formula $\Diamond K_\alpha[\alpha]\neg\varphi$—which is strictly stronger than $K_\alpha\Diamond\neg[\alpha]\varphi$—is used as the negative condition of motivational-active responsibility; and formula $K_\alpha\neg[\alpha]\neg\varphi \wedge I_\alpha\neg[\alpha]\neg\varphi \wedge \Diamond K_\alpha[\alpha]\neg\varphi$—which is strictly stronger than $I_\alpha\neg[\alpha]\neg\varphi \wedge K_\alpha\Diamond[\alpha]\neg\varphi$—is used as the negative conditions of motivational-passive responsibility. Therefore, α was motivational-active responsible for φ at $\langle m,h\rangle$ only if it was possible for α to knowingly prevent φ at $\langle m,h\rangle$—that is, only if $\Diamond K_\alpha[\alpha]\neg\varphi$ holds at the index. Otherwise, someone could claim that α knew that it was impossible to knowingly prevent φ (formula $\Box\neg K_\alpha[\alpha]\neg\varphi \to K_\alpha\Box\neg K_\alpha[\alpha]\neg\varphi$ is valid), and that such knowledge somehow led α to knowingly and intentionally bring about φ. Similar arguments apply to motivational-passive responsibility: the negative conditions imply a policy of leniency on blameworthy agents.

As mentioned before, the negative conditions in Table 6.1 were set so that the presentation of modes would be simpler. However, the reader is encouraged to explore different options for characterizing the sub-categories. Indeed, the logic is flexible enough to make diverse proposals in this respect, where each proposal could be accordingly labelled within a spectrum of leniency-to-strictness on agents.

As a final note to the discussion on the sub-categories of responsibility, let me address an important observation concerning the relations between these sub-categories:

Observation 6.4. *Let $\mathcal{M}$ be a* kiobt-*model. For all $\alpha \in Ags$, index $\langle m,h\rangle$, and φ of $\mathcal{L}_R$, the following items hold:*

[16] This validity is implied by schema (D) for I_α and Observation 5.9 a (p. 244).

1. (a) *If α was informational-active, resp. informational-passive, responsible for φ at $\langle m, h\rangle$, then α was causal-active, resp. causal-passive, responsible for φ at $\langle m, h\rangle$; the converse is not true.*

   The implication follows from the validity of $\Diamond K_\alpha\psi \rightarrow K_\alpha\Diamond\psi$ and of schema (T) for K_α. An example of the fact that the converse does not hold is given by Chapter 5's Example 5.7 (p. 231): at $\langle m_3, h_3\rangle$ *doctor* was causal-active and causal-passive responsible for the patient's death, but she was neither informational-active nor informational-passive responsible for it.

   (b) *If α was motivational-active, resp. motivational-passive, responsible for φ at $\langle m, h\rangle$, then α was informational-active, resp. informational-passive, responsible for φ at $\langle m, h\rangle$; the converse is not true.*

   The implication follows from the definitions of informational and motivational responsibility. An example of the fact that the converse does not hold is given by Chapter 5's Figure 5.4 (p. 239): at $\langle m_1, h_1\rangle$ *driver* was informational-active and informational-passive responsible for the officer's death, but she was neither motivational-active nor motivational-passive responsible for it.

2. *For all three categories, the active form of responsibility with respect to φ implies the passive form.*

   For causal responsibility, the result follows from the validity of schema (T) for $[\alpha]$. For informational responsibility, it follows from the validity of schemata (T) for K_α and $[\alpha]$, the validity of schema (D) for $[\alpha]$, Necessitation for K_α, and the validity of schema (K) for K_α. For informational responsibility, it follows from the validity of schema (T) for $[\alpha]$, the validity of schema (D) for $[\alpha]$, Necessitation for I_α, and the validity of schema (K) for I_α.

After this discussion on the sub-categories of responsibility, we are ready to introduce the full characterizations of the modes of responsibility. As has often been mentioned in this chapter, this is done by collating the sub-categories against the backdrop of specific deontic contexts.

6.3.3 Formalization of Modes of Responsibility

In Section 6.2 I explained why blame-or-praise assignment is one of the main reasons for considering the relation '*... is responsible for...*' relevant at all. Furthermore, the examples given in said section illustrate the intuition that agents' obligations—or ought-to-do's—are deeply connected with systematic blame-or-praise assignment. To be precise, obligations provide the deontic contexts of responsibility, that determine degrees of praiseworthiness/blameworthiness for instances of the notion. Aiming for clarity in the presentation of my proposal for a system of such degrees, I will anchor the modes of responsibility to the different deontic contexts, where these contexts are represented with conjunctions of the deontic modalities of *IEAUST*.[17]

Let $\mathcal{M}$ be a *kiobt*-model. Take $\alpha \in Ags$, and let φ be a formula of $\mathcal{L}_{\mathsf{R}}$. For each index $\langle m, h\rangle$, there are 4 main possibilities for conjunctions of deontic modalities holding at $\langle m, h\rangle$, according to whether $\Delta\varphi$ or $\neg\Delta\varphi$ is satisfied at the index, where $\Delta \in \left\{\odot_\alpha, \odot_\alpha^S\right\}$. I refer to any such conjunction as a *deontic context for α's responsibility with respect to φ at $\langle m, h\rangle$*. Thus, these contexts render 4 main levels of praiseworthiness, resp. blameworthiness, under the premise that bringing about φ is praiseworthy and refraining from bringing about φ is blameworthy. I use numbers 1–4 to refer to these levels, so that *Level* 1 corresponds the highest level of praiseworthiness, resp. blameworthiness, and *Level* 4 corresponds to the lowest level.

Level 1: when deontic context $\odot_\alpha\varphi \land \odot_\alpha^S\varphi$ holds at $\langle m, h\rangle$.

Compared with the other levels below—and *generally speaking*—in this level bringing about φ comes with the highest degree of praiseworthiness, and refraining from bringing about φ comes with the highest degree of blameworthiness.[18] Observe that $\mathcal{M}, \langle m, h\rangle \models \odot_\alpha\varphi \land \odot_\alpha^S\varphi$ iff at $\langle m, h\rangle$ α objectively and subjectively ought to have seen to it that φ. In other words, at $\langle m, h\rangle$ α was compelled on both deontic accounts to bring about φ. A good example of this level of compulsion was given in Section 6.2, where you are objectively and subjectively obligated to stop your car at a crossing walk, to let school children pass. If you stop your car, then you are abiding by the law, and the legal system will praise you for it. If you do not stop your car, then the legal system will investigate how much you are to blame for any tragic outcome that ensues from your action.

For deontic context $\odot_\alpha\varphi \land \odot_\alpha^S\varphi$, the *basic modes of α's active responsibility with respect to φ at $\langle m, h\rangle$* are displayed in Tables 6.2 and 6.3, where the latter is the transcription of the former in terms of formulas of $\mathcal{L}_{\mathsf{R}}$. In both tables, the columns are indexed by the attribution-attitudes of praiseworthiness/blameworthiness (***Att.***), and the rows are indexed by degrees for such attitudes (***Deg.***).

[17] As mentioned before, the deontic contexts account for constituent (3) in any given mode.

[18] As established in Section 6.2, this does not imply that the degree of praiseworthiness attributed to bringing about φ is exactly the opposite of the degree of blameworthiness attributed to refraining from bringing about φ. Sometimes complying with an obligation does not elicit a level of praiseworthiness that is comparable to the level of blameworthiness that failing to comply elicits (recall the example involving stopping your car at a crossing walk). Indeed, these nuances depend on what φ actually means.

Att. ***Deg.***	**Praiseworthiness**	**Blameworthiness**
Low_A	Causal-active for φ ✓ Infor.-active for φ ✗ Motiv.-active for φ ✗	Causal-active for $\neg\varphi$ ✓ Infor.-active for $\neg\varphi$ ✗ Motiv.-active for $\neg\varphi$ ✗
$Middle_A$	Causal-active for φ ✓ Infor.-active for φ ✓ Motiv.-active for φ ✗	Causal-active for $\neg\varphi$ ✓ Infor.-active for $\neg\varphi$ ✓ Motiv.-active for $\neg\varphi$ ✗
$High_A$	Causal-active for φ ✓ Infor.-active for φ ✓ Motiv.-active for φ ✓	Causal-active for $\neg\varphi$ ✓ Infor.-active for $\neg\varphi$ ✓ Motiv.-active for $\neg\varphi$ ✓

Table 6.2: *Modes of α's active responsibility with respect to φ.*

Att. ***Deg.***	**Praiseworthiness**	**Blameworthiness**
Low_A	$([\alpha]\varphi \wedge \Diamond[\alpha]\neg\varphi) \wedge$ $(\neg K_\alpha[\alpha]\varphi \vee \neg\Diamond K_\alpha[\alpha]\neg\varphi) \wedge$ $(\neg I_\alpha[\alpha]\varphi \vee \neg\Diamond K_\alpha[\alpha]\neg\varphi)$	$([\alpha]\neg\varphi \wedge \Diamond[\alpha]\varphi) \wedge$ $(\neg K_\alpha[\alpha]\neg\varphi \vee \neg\Diamond K_\alpha[\alpha]\varphi) \wedge$ $(\neg I_\alpha[\alpha]\neg\varphi \vee \neg\Diamond K_\alpha[\alpha]\varphi)$
$Middle_A$	$(K_\alpha[\alpha]\varphi \wedge \Diamond K_\alpha[\alpha]\neg\varphi) \wedge$ $\neg I_\alpha[\alpha]\varphi$	$(K_\alpha[\alpha]\neg\varphi \wedge \Diamond K_\alpha[\alpha]\varphi) \wedge$ $\neg I_\alpha[\alpha]\neg\varphi$
$High_A$	$(K_\alpha[\alpha]\varphi \wedge \Diamond K_\alpha[\alpha]\neg\varphi) \wedge$ $I_\alpha[\alpha]\varphi$	$(K_\alpha[\alpha]\neg\varphi \wedge \Diamond K_\alpha[\alpha]\varphi) \wedge$ $I_\alpha[\alpha]\neg\varphi$

Table 6.3: *Modes of α's active responsibility with respect to φ.*

Therefore, Tables 6.2 and 6.3 include those basic modes—of α's responsibility with respect to φ at $\langle m,h\rangle$—whose constituent (3) is deontic context $\odot_\alpha\varphi \wedge \odot^S_\alpha\varphi$. To illustrate these degrees—for context $\odot_\alpha\varphi \wedge \odot^S_\alpha\varphi$—let me explicitly discuss the cells of Tables 6.2 and 6.3.

Assume that $\mathcal{M},\langle m,h\rangle \models \odot_\alpha\varphi \wedge \odot^S_\alpha\varphi$. As for degrees of **praiseworthiness**, Low_A applies to the mode characterized with formula $([\alpha]\varphi \wedge \Diamond[\alpha]\neg\varphi) \wedge (\neg K_\alpha[\alpha]\varphi \vee \neg\Diamond K_\alpha[\alpha]\neg\varphi) \wedge (\neg I_\alpha[\alpha]\varphi \vee \neg\Diamond K_\alpha[\alpha]\neg\varphi)$. If this formula holds at $\langle m,h\rangle$, then α was causal-active responsible for φ at $\langle m,h\rangle$, but α was not informational-active nor motivational-active responsible. In other words, α brought about φ unknowingly and unintentionally, and it was possible for α to knowingly prevent φ. Now, to illustrate why this mode gets a low degree of praiseworthiness, consider the example of stopping your car at a crossing walk. Let s stand for the proposition 'your car is stopped.' Then formula $\odot_{you}s \wedge \odot^{S}_{you}s$ holds. Consider a version of this example where you indeed stop your car, but only as a lucky coincidence. Knowing that it would not be deontically ideal, you were actually trying to speed up—just for the selfish reason of being fast—and you mistook the brake pedal for the accelerator. Thus, formula $[you]s \wedge \neg K_{you}[you]s \wedge \neg I_{you}[you]s$ holds: *at the implied actual index you stopped your car but did so unknowingly and unintentionally.* Whoever happens to know this background information would likely praise you very mildly, since you only complied with your obligation by mere chance and without having any intention of doing so.[19]

It is important to mention that if $\mathcal{M}, \langle m,h\rangle \models \odot_{\alpha}\varphi \wedge \odot^{S}_{\alpha}\varphi$ then $\mathcal{M}, \langle m,h\rangle \models \Diamond K_{\alpha}[\alpha]\varphi$. Therefore, for deontic context $\odot_{\alpha}\varphi \wedge \odot^{S}_{\alpha}\varphi$, it must be the case that α should be able to knowingly see to it that φ. This is a consequence of the validity of $\odot^{S}_{\alpha}\varphi \rightarrow \Diamond K_{\alpha}[\alpha]\varphi$ (a subjective version of Kant's directive of *ought implies can*, discussed in item 9 in the list of *EAUST*'s logic-based properties, Chapter 4, p. 174). In the example above, I take it that it is possible for you to knowingly stop the car, since you can use the hand-brake instead of the brake pedal.

For the praiseworthy mode tagged $Middle_A$, formula $(K_{\alpha}[\alpha]\varphi \wedge \Diamond K_{\alpha}[\alpha]\neg\varphi) \wedge \neg I_{\alpha}[\alpha]\varphi$ holds at $\langle m,h\rangle$. Thus, α was causal-active and informational-active responsible for φ at $\langle m,h\rangle$, but α was not motivational-active responsible. In other words, α knowingly brought about φ, did so unintentionally, and it was possible for α to knowingly prevent φ. To illustrate why this mode gets a middle degree of praiseworthiness, recall the basketball example from Section 6.2 (p. 270). In this example, you were a 'ball hog' that did not p-d intend to pass the ball to the strongest player in the team for the last play of the game. You wanted to take all the credit by scoring on your own, and you only passed the ball because the head coach had previously told you that otherwise you would be fired. If p stands for the proposition 'the ball is passed to the strongest player in your team,' then formulas $\odot_{you}p \wedge \odot^{S}_{you}p$, $K_{you}[you]p$, and $\neg I_{you}[you]p$ hold: *at the implied actual*

[19] Moreover, the statement of this example implies that formulas $\neg K_{you}\neg[you]\neg s$ and $I_{you}[you]\neg s$ also hold. Therefore, you did not knowingly refrain from not stopping your car, and you p-d intended to see to it that your car did not stop. Thus, you were *not* informational-passive *nor* motivational-passive responsible for stopping the car, and you could have been! Therefore, your degree of praiseworthiness might still be less than the one already described. This situation highlights that the present framework allows the use of many combinations of modalities to describe nuances in responsibility and blame-or-praise assignment (see the discussion after the presentation of modes of passive responsibility on p. 287).

index you were compelled to pass the ball on both deontic accounts, and you knowingly but unintentionally passed the ball. Since it was possible for you to knowingly not pass the ball, then you are causal-active and informational-active responsible for passing, but you are not motivational-active responsible. Thus, and as mentioned before, in this case your teammates would not likely praise your action very much. Still, this mode generally receives more praise than the one labelled Low_A, since to knowingly comply with an obligation typically involves certain skills; in other words, the agent did not comply by mere chance.

For the praiseworthy mode tagged $High_A$, formula $(K_\alpha[\alpha]\varphi \wedge \Diamond K_\alpha[\alpha]\neg\varphi) \wedge I_\alpha[\alpha]\varphi$ holds at $\langle m, h\rangle$. Thus, α was causal-active, informational-active, and motivational-active responsible for φ at $\langle m, h\rangle$. In other words, α knowingly and intentionally brought about φ, while it was possible for α to knowingly prevent φ. To illustrate why this mode gets the highest degree of praiseworthiness, consider a different version of the basketball example above, where the strongest player in the team is you. Here, you knowingly and intentionally take the last shot—as part of a play that you and your teammates ran almost to perfection, just as designed by the head coach. If b stands for the proposition 'the ball is shot,' then formulas $\odot_{you}b \wedge \odot^{S}_{you}b$, $I_{you}[you]b$, and $K_{you}[you]b$ hold: *at the implied actual index you knowingly and intentionally fulfilled your obligation of executing the last play to the best of your ability*. Regardless of whether the shot goes in, it is likely that everyone will highly praise you for your action.

As for degrees of **blameworthiness**, Low_A applies to the mode characterized with $([\alpha]\neg\varphi \wedge \Diamond[\alpha]\varphi) \wedge (\neg K_\alpha[\alpha]\neg\varphi \vee \neg\Diamond K_\alpha[\alpha]\varphi) \wedge (\neg I_\alpha[\alpha]\neg\varphi \vee \neg\Diamond K_\alpha[\alpha]\varphi)$. If this formula holds at $\langle m, h\rangle$, then α was causal-active responsible for $\neg\varphi$ at $\langle m, h\rangle$, but α was not informational-active nor motivational-active responsible. To illustrate why this mode gets a low degree of blameworthiness, consider the version of the driver example that was mentioned on p. 269. You are driving your car, approaching a crossing walk where a traffic officer is holding up a stop-sign so that a group of school children can cross the road. Thus, formula $\odot_{you}s \wedge \odot^{S}_{you}s$ holds, where s stands for the proposition 'your car is stopped.' This time, a terrorist had previously hacked the wiring and the computer of your car. Right before you reached the crossing walk, the terrorist—who knew that you were approaching the crossing walk because he had also installed camera on the front hood—remotely changed the settings of your car so that the brake pedal turned into the accelerator and vice versa. At the crossing walk, you stepped on the brake pedal to stop the car, but the car sped up, and a tragedy occurred. Thus, formula $[you]\neg s \wedge \neg K_{you}[you]\neg s \wedge \neg I_{you}[you]\neg s$ holds: *at the implied actual index you* *kept on driving; however, you did not knowingly nor intentionally keep on driving.*[20] In this case, if you are taken to court for your involvement in the tragedy, and if your lawyers demonstrate that you are neither informational-active nor motivational-active responsible for keeping on driving, then it is likely that the jury will either absolve you or blame you very mildly, because the really blameworthy agent is the terrorist.

For the blameworthy mode tagged $Middle_A$, formula $[\alpha]\neg\varphi \wedge \neg K_\alpha[\alpha]\neg\varphi \wedge (I_\alpha[\alpha]\neg\varphi \wedge \Diamond K_\alpha[\alpha]\varphi)$ holds at $\langle m,h\rangle$. Thus, α was causal-active and informational-active responsible for $\neg\varphi$ at $\langle m,h\rangle$, but α was not motivational-active responsible. To illustrate why this mode gets a middle degree of blameworthiness, consider another version of the driver example (mentioned on p. 269). In this case, there is a terrorist inside your car, holding your family at gun point. This terrorist has threatened to shoot your family if you do not keep on driving. Making a difficult decision, you keep on driving. Thus, formula $[you]\neg s \wedge K_{you}[you]\neg s \wedge \neg I_{you}[you]\neg s$ holds: *at the implied actual index you knowingly but unintentionally kept on driving.* Thus, if you are taken to court, and if your lawyers demonstrate that you are not motivational-active responsible for keeping on driving, it is likely that the jury will either absolve you or blame you very mildly, since the blameworthy agent is the terrorist. Still, this mode generally receives more blame than the one labelled Low_A, since it implies that the agent knowingly brought about the undesirable state of affairs.

For the blameworthy mode tagged $High_A$, formula $(K_\alpha[\alpha]\neg\varphi \wedge \Diamond K_\alpha[\alpha]\varphi) \wedge I_\alpha[\alpha]\neg\varphi$ holds at $\langle m,h\rangle$. Thus, α was causal-active, informational-active, and motivational-active responsible for $\neg\varphi$ at $\langle m,h\rangle$. To illustrate why this mode gets the highest degree of blameworthiness, consider the first driver example in Section 6.2 (p. 269), itself based on Chapter 5's Example 5.6 (see Figure 5.1 on p. 231). Here, you approach the crossing walk in your car, and you knowingly and intentionally keep on driving while a traffic officer is holding a stop sign so that school children can cross the road. Thus, formula $K_{you}[you]\neg s \wedge I_{you}[you]\neg s$ holds: *at the implied actual index you knowingly and intentionally kept on driving.* In this case, it would be difficult to say that you are not highly blameworthy for whatever terrible outcome that occurs.

[20] Recall that the subjective version of Kant's directive of *ought implies can* yields that it must be possible for you to knowingly stop your car (formula $\Diamond K_{you}[you]s$ must hold). As before, I take it that it is possible for you to knowingly stop the car, since you can use the hand-brake instead of the brake pedal.

Observe that Tables 6.2 and 6.3 concern the active form of responsibility. As for passive responsibility, the basic modes of α's passive responsibility with respect to φ are included in Tables 6.4 and 6.5, where the latter is the transcription of the former in terms of formulas of $\mathcal{L}_{\mathsf{R}}$.

Att. ***Deg.***	**Praiseworthiness**	**Blameworthiness**
Low_P	Causal-passive for φ ✓ Infor.-passive for φ ✗ Motiv.-passive for φ ✗	Causal-passive for $\neg\varphi$ ✓ Infor.-passive for $\neg\varphi$ ✗ Motiv.-passive for $\neg\varphi$ ✗
Middle_P	Causal-passive for φ ✓ Infor.-passive for φ ✓ Motiv.-passive for φ ✗	Causal-passive for $\neg\varphi$ ✓ Infor.-passive for $\neg\varphi$ ✓ Motiv.-passive for $\neg\varphi$ ✗
High_P	Causal-passive for φ ✓ Infor.-passive for φ ✓ Motiv.-passive for φ ✓	Causal-passive for $\neg\varphi$ ✓ Infor.-passive for $\neg\varphi$ ✓ Motiv.-passive for $\neg\varphi$ ✓

Table 6.4: *Modes of α's passive responsibility with respect to φ.*

Att. ***Deg.***	**Praiseworthiness**	**Blameworthiness**
Low_P	$(\varphi \wedge \Diamond[\alpha]\neg\varphi) \wedge$ $(\neg K_\alpha\neg[\alpha]\neg\varphi \vee \neg\Diamond K_\alpha[\alpha]\neg\varphi) \wedge$ $(\neg I_\alpha\neg[\alpha]\neg\varphi \vee \neg\Diamond K_\alpha[\alpha]\neg\varphi)$	$(\neg\varphi \wedge \Diamond[\alpha]\varphi) \wedge$ $(\neg K_\alpha\neg[\alpha]\varphi \vee \neg\Diamond K_\alpha[\alpha]\varphi) \wedge$ $(\neg I_\alpha\neg[\alpha]\varphi \vee \neg\Diamond K_\alpha[\alpha]\varphi)$
Middle_P	$\varphi\wedge$ $(K_\alpha\neg[\alpha]\neg\varphi \wedge \Diamond K_\alpha[\alpha]\neg\varphi) \wedge$ $\neg I_\alpha\neg[\alpha]\neg\varphi$	$\neg\varphi\wedge$ $(K_\alpha\neg[\alpha]\varphi \wedge \Diamond K_\alpha[\alpha]\varphi) \wedge$ $\neg I_\alpha\neg[\alpha]\varphi$
$\mathit{Highest}_P$	$\varphi\wedge$ $(K_\alpha\neg[\alpha]\neg\varphi \wedge \Diamond K_\alpha[\alpha]\neg\varphi) \wedge$ $I_\alpha\neg[\alpha]\neg\varphi$	$\neg\varphi\wedge$ $(K_\alpha\neg[\alpha]\varphi \wedge \Diamond K_\alpha[\alpha]\varphi) \wedge$ $I_\alpha\neg[\alpha]\varphi$

Table 6.5: *Modes of α's passive responsibility with respect to φ.*

An interesting question to address, then, is how the degrees of praiseworthiness/blameworthiness for the modes of passive responsibility compare with those for active responsibility. According to Observation 6.4 2, the active form implies the passive form in all categories. Thus, it is reasonable to assume that, for any tag $T \in \{\mathit{Low}, \mathit{Middle}, \mathit{High}\}$, the basic mode tagged T_A comes with a bit more praiseworthiness/blameworthiness than the mode tagged T_P.

Furthermore, the active/passive dichotomy makes the gradation of praiseworthiness/blameworthiness still more complex. To clarify, consider Table 6.4. For any basic mode X in such a table, consider all the sub-modes that result from substituting the term 'active' for the term 'passive' in some check-marked category in X. It is reasonable to assume that the more categories are taken to be active, the higher the degree of praiseworthiness/blameworthiness for the sub-mode should be. However, it is relatively unclear how to compare the degrees for such sub-modes with those of modes with tags involving more praiseworthiness/blameworthiness than that of X. In particular, for tags $T, S \in \{\mathit{Low}, \mathit{Middle}, \mathit{Highest}\}$, it is unclear how to compare T_A and S_P when S is higher in praiseworthiness/blameworthiness than T. To illustrate this conundrum, consider Figure 6.1.

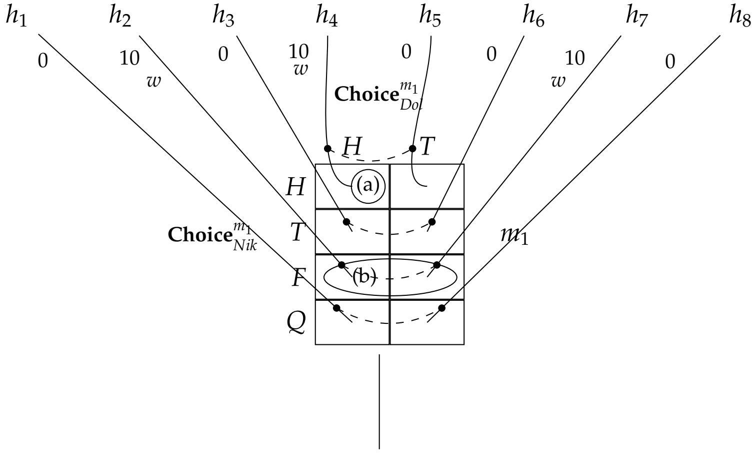


Figure 6.1: *Matching kopecks pt. 2.*

Figure 6.1 depicts another game that *Nikolai* and *Dolokhov* are playing at a gambling house. Here, *Nikolai* is holding a kopeck in his hand, and at m_1 he is faced with four options: placing his kopeck on a table heads up (labelled by H), placing his kopeck tails up (T), forfeiting the bet (F), or quitting the game and leaving the gambling house (Q). At the same moment, *Dolokhov* must place his kopeck on the table, either heads up (H) or tails up (T). No previous communication between the two players is allowed so as to prevent them from aligning their choices beforehand. If *Nikolai* and *Dolokhov* both place their kopecks heads up or both place their kopecks tails up, the house gives them 10 roubles; if *Nikolai* forfeits the bet, the house also pays 10 roubles; finally, if *Nikolai* quits the game, the house pays nothing.

Let us focus on *Nikolai*'s situation. *Nikolai*'s epistemic states at m_1 are represented in the diagram with the indistinguishability relation given by dashed lines. Thus, if w stands for the proposition 'Nikolai and Dolokhov win,' then formulas $\odot_{Nik} w \land \odot^{S}_{Nik} w$, $\Diamond K_{Nik}[Nik]w$, $\Diamond K_{Nik}[Nik]\neg w$, and $\Diamond K_{Nik}\neg[Nik]w$ all hold at every index based on m_1. Consider, then, the following two modes:

(a) Suppose that *Nikolai* p-d intended to win but he could not bring himself to forfeit the bet, thinking that this would be a 'cowardly' option. Thus, he chose to gamble and placed his kopeck heads up on the table, hoping to win. This is modelled in the diagram with the p-d intention embodied by the circle enclosing tag '(a).' Suppose that they nonetheless lost. This means that *Dolokhov* played tails, so that the actual index is $\langle m_1, h_5\rangle$. Observe, then, that $\mathcal{M}, \langle m_1, h_5\rangle \models \neg w \land I_{Nik}\neg[Nik]w \land K_{Nik}\neg[Nik]w$: *at* $\langle m_1, h_5\rangle$ Nikolai *was motivational-passive for losing*. According to Tables 6.4 and 6.5, the mode of *Nikolai*'s responsibility with respect to w at $\langle m_1, h_5\rangle$ is tagged blameworthy $High_P$.

(b) Suppose that *Nikolai* p-d intended to win by forfeiting the bet and that he was forced by someone in the gambling house to quit the game. This is modelled in the diagram with the p-d intention embodied by the ellipse enclosing tag '(b).' Without loss of generality, then, assume that the actual index is $\langle m_1, h_1\rangle$, where $\mathcal{M}, \langle m_1, h_1\rangle \models K_{Nik}[Nik]\neg w \land I_{Nik}[Nik]w$: *at* $\langle m_1, h_1\rangle$ Nikolai *is informational-active responsible for losing, but he is not motivational-passive responsible*. According to Tables 6.2 and 6.3, the mode of *Nikolai*'s responsibility with respect to w at $\langle m_1, h_1\rangle$ is tagged blameworthy $Middle_A$.

How should one compare the blameworthiness of the mode in case (a) with that of the mode in case (b)? In (a) *Nikolai* knowingly and intentionally refrained from winning, but he chose an action that could have led to winning. In contrast, in (b) *Nikolai* knowingly—but unintentionally—lost, choosing an action for which there was no possibility of winning. Thus, even if the tag for the mode in case (a) implies that its blameworthiness is higher than that of the mode in case (b), some people would definitely say otherwise. At this point, I shy away from settling this dilemma, because such a thing lies beyond the scope of my discussion. However, it is important to mention that *IEAUST* is flexible enough to include many nuances in the gradation of praiseworthiness/blameworthiness, that can be set as the reader sees fit.

Along the same lines, observe that the validity of $\odot^S_\alpha\varphi \to K_\alpha\Box \odot^S_\alpha \varphi$ implies that if deontic context $\odot_\alpha\varphi \wedge \odot^S_\alpha\varphi$ holds at $\langle m, h\rangle$ then at this index α knew *ex ante* that it subjectively ought to have seen to it that φ.[21] As such, it can be said that α was conscious of its subjective obligation to bring about φ. In my view, this implies that the obligation was all the more compelling. In contrast, it is neither necessarily true that (a) at $\langle m, h\rangle$ α knew *ex ante* that it objectively ought to have seen to it that φ nor necessarily true that (b) at $\langle m, h\rangle$ α knew *ex interim* about this objective obligation. Indeed, formulas $\neg\Box K_\alpha \odot_\alpha \varphi$ and $\neg K_\alpha \odot_\alpha \varphi$ can hold at $\langle m, h\rangle$. Therefore, for $\psi \in \{\Box K_\alpha \odot_\alpha \varphi, K_\alpha \odot_\alpha \varphi\}$, the conjunctions $\left(\odot_\alpha\varphi \wedge \odot^S_\alpha\varphi\right) \wedge \psi$ and $\left(\odot_\alpha\varphi \wedge \odot^S_\alpha\varphi\right)\wedge\neg\psi$ imply different sub-cases of the main deontic context $\odot_\alpha\varphi\wedge\odot^S_\alpha\varphi$, which lead to more nuances in the gradation of modes of responsibility. In fact, the same claim of existence of sub-cases applies to *Levels* 2–3 below, substituting $\left(\odot_\alpha\varphi \wedge \odot^S_\alpha\varphi\right)$ for any of the deontic contexts of such levels.

Level 2: when deontic context $\neg \odot_\alpha \varphi \wedge \odot^S_\alpha\varphi$ holds at $\langle m, h\rangle$.

Observe that $\mathcal{M}, \langle m, h\rangle \models \neg \odot_\alpha \varphi \wedge \odot^S_\alpha\varphi$ iff at $\langle m, h\rangle$ α subjectively ought to have seen to it that φ, but α did not objectively ought to have seen to it that φ. Now, $\neg \odot_\alpha \varphi$ holds at $\langle m, h\rangle$ iff φ is not an effect of all objectively optimal actions—for α—at m. More precisely, there must exist $L \in \mathbf{Optimal}^m_\alpha$ such that $L \nsubseteq |\varphi|^m$.[22] By the semantics for the two deontic modalities (Definition 6.3), this means that L is just as good, in the objective dominance ordering $\preceq$, as every action in $\mathbf{SOptimal}^m_\alpha$—where, for all L' in $\mathbf{SOptimal}^m_\alpha$, $[L']^{m'}_\alpha \subseteq |\varphi|^m$ for every m' such that $m \sim_\alpha m'$. The existence of objectively optimal actions that do not support φ justifies the claims that—*generally speaking*—bringing about φ in this level comes with less praiseworthiness than in *Level 1* and refraining from bringing about φ in this level comes with less blameworthiness than in *Level 1*.

To illustrate the present deontic context, consider another example of *Nikolai* and *Dolokhov*. This time, the set-up is the one from Chapter 4's Figure 4.7 (p. 167): if *Nikolai* bets and chooses correctly, or if he forfeits the bet, *Nikolai* and *Dolokhov* win 10 roubles. If he chooses incorrectly, they win nothing. Observe that $\mathbf{SOptimal}^{m_2}_{Nik} = \{N_3\}$, and $\mathbf{SOptimal}^{m_3}_{Nik} = \{N_6\}$. Recall that f stands for the proposition 'Nikolai has forfeited the bet.' According to Definition 6.3, $\mathcal{M}, \langle m_i, h\rangle \models \neg\odot_{Nik} f \wedge \odot^S_{Nik} f$ for all $i \in \{2, 3\}$ and $h \in H_{m_i}$: *at all indices based on* m_2 *and* m_3 *Nikolai did not objectively ought to have forfeited the bet, but he subjectively ought to*

[21] This chapter adopts the same conventions, with respect to the topics that are relevant in *EST* (*knowingly doing, epistemic sense of ability, knowledge across the stages of information disclosure,* and *uniformity*), as Chapter 4 (see Subsection 4.2.2). For a discussion on the validity of $\odot^S_\alpha\varphi \to K_\alpha\Box \odot^S_\alpha \varphi$, see item 11 in the list of logic-based properties for Chapter 4's *EAUST* (Subsection 4.5.1, p. 175).

[22] Recall from Definition 6.3 that I write $|\varphi|^m$ to refer to the set $\{h \in H_m; \mathcal{M}, \langle m, h\rangle \models \varphi\}$.

have forfeited the bet. Observe that at m_2 there is an objectively optimal action—for *Nikolai*—that does not imply forfeiting. Namely, the action N_1 of choosing heads is such that $N_1 \in \mathbf{Optimal}^{m_2}_{Nik}$ and such that $N_1 \not\subseteq \left|f\right|^{m_2}$. Similarly, at m_3 the action N_5 of choosing tails is such that $N_5 \in \mathbf{Optimal}^{m_3}_{Nik}$ and such that $N_5 \not\subseteq \left|f\right|^{m_3}$.

The basic modes of responsibility associated with deontic context $\neg \odot_\alpha \varphi \wedge \odot^S_\alpha \varphi$ are also given in Tables 6.2, 6.3, 6.4, and 6.5, with the same degrees of praiseworthiness/blameworthiness as displayed there, but relative to the present deontic context. When comparing this level with *Level 1*, one must be careful. It is not that the mode with lowest praiseworthiness, resp. blameworthiness, of *Level 1* has higher praiseworthiness, resp. blameworthiness, than the mode tagged 'highest' in *Level 2*. Rather, for a fixed conjunction of formulas characterizing a combination of sub-categories of α's responsibility with respect to φ at $\langle m, h\rangle$, the mode that results from coupling this conjunction with the deontic context of *Level 1* gets a higher degree of praiseworthiness, resp. blameworthiness, than the mode that results from coupling this conjunction with the deontic context of *Level 2*. In other words, suppose that two modes x and y, of α's responsibility with respect to φ at $\langle m, h\rangle$, share constituents (1) and (2) (where these constituents refer to the sub-categories of responsibility according to the active and passive forms). If constituent (3) of mode x is the deontic context of *Level 1* and constituent (3) of mode y is the deontic context of *Level 2*, then x gets a higher degree of praiseworthiness, resp. blameworthiness, than y. For instance, suppose that α was informational-active and motivational-active responsible for φ at $\langle m, h\rangle$. Then the praise that α deserves if the deontic context is $\odot_\alpha \varphi \wedge \odot^S_\alpha \varphi$ should be higher than the one deserved if the deontic context were $\neg \odot_\alpha \varphi \wedge \odot^S_\alpha \varphi$. *It is important to emphasize that this is the criterion of comparison between all four levels.*

The intuition that this level involves less blameworthiness than *Level 1*—for a fixed responsibility-related formula—can be illustrated in terms of α's excusability: if α engages in a blameworthy mode in this level, then α can be excused for not having brought about φ if one claims that $\neg[\alpha]\varphi$ resulted from having chosen one of the objectively optimal actions that did not enforce φ. This would be a rather weak excuse, though, because the properties of subjective ought-to-do's imply that (a) it was possible for α to have knowingly brought about φ, and that (b) α knew that it subjectively ought to have seen to it that φ (recall that formulas $\odot^S_\alpha \varphi \to \Diamond K_\alpha[\alpha]\varphi$ and $\odot^S_\alpha \varphi \to K_\alpha \Box \odot^S_\alpha \varphi$ are valid). For instance, suppose that in the example above *Nikolai* chooses to bet heads instead of forfeiting the bet, and that he loses. Therefore, in Figure 4.7 *Nikolai* chose action N_4, and the actual index is $\langle m_3, h_4\rangle$. In principle, *Nikolai* can be excused for not forfeiting by claiming that his choice (N_4) was, in his view, epistemically equivalent to an objectively optimal action where to bet heads led to his winning (N_1). Since he knew that by forfeiting the bet he would have knowingly won, and since he knew that it was possible for him to knowingly forfeit, then, for such an excuse to actually work, he would probably have to justify it on the grounds of some personal intention, belief, or long-term strategy, for instance.

Similar to what was said for the deontic context of the previous level, for $\psi \in \{\Box K_\alpha \neg \odot_\alpha \varphi, K_\alpha \neg \odot_\alpha \varphi\}$, the conjunctions $\left(\neg \odot_\alpha \varphi \wedge \odot^S_\alpha \varphi\right) \wedge \psi$ and $\left(\neg \odot_\alpha \varphi \wedge \odot^S_\alpha \varphi\right) \wedge \neg\psi$ present sub-cases of the main deontic context $\neg \odot_\alpha \varphi \wedge \odot^S_\alpha \varphi$.

Level 3: when deontic context $\odot_\alpha\varphi \wedge \neg \odot^S_\alpha \varphi$ holds at $\langle m, h\rangle$.

The basic modes associated with the deontic context of this level are once again displayed in Tables 6.2, 6.3, 6.4, and 6.5.

Observe that $\mathcal{M}, \langle m, h\rangle \models \odot_\alpha\varphi \wedge \neg \odot^S_\alpha \varphi$ iff at $\langle m, h\rangle$ α objectively ought to have seen to it that φ, but α did not subjectively ought to have seen to it that φ. Therefore, there exists $L \in \mathbf{SOptimal}^m_\alpha$ such that $[L]^{m'}_\alpha \not\subseteq |\varphi|^m$ for some m' such that $m \sim_\alpha m'$. This means that L is just as good, in the subjective dominance ordering $\leq_s$, as every action in $\mathbf{Optimal}^m_\alpha$—where, for all $L' \in \mathbf{Optimal}^m_\alpha$, $L' \subseteq |\varphi|^m$. An example of this situation can be found in the *Miners Paradox* discussed at the beginning of Chapter 4 (p. 139). To clarify, consider the *kiobt*-model depicted in Figure 6.2.

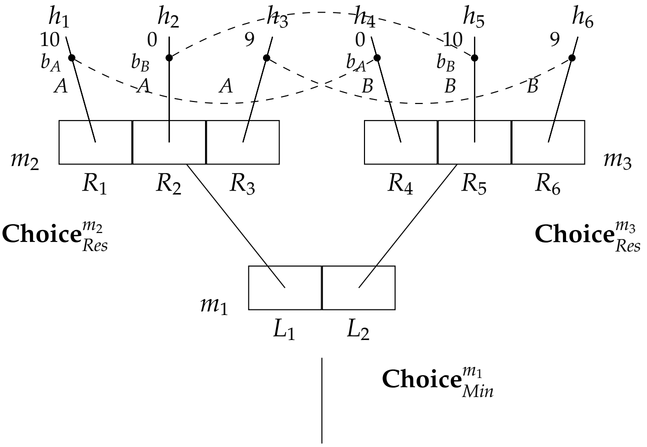


Figure 6.2: *Miners paradox.*

In Figure 6.2, L_1 represents the choice, available to the miners, of going into shaft A, and L_2 represents the choice of going into shaft B; R_1 and R_4 represent the choices, available to the rescuers, of blocking shaft A; R_2 and R_5 represent the choices of blocking shaft B; and R_3 and R_6 represent the choices of refraining from blocking any shaft. Observe that $\mathbf{Optimal}^{m_2}_{Res} = \{R_1\}$, that $\mathbf{Optimal}^{m_3}_{Res} = \{R_5\}$, that $\mathbf{SOptimal}^{m_2}_{Res} = \{R_1, R_2, R_3\}$, and that $\mathbf{SOptimal}^{m_3}_{Res} = \{R_4, R_5, R_6\}$. Let A stand for the proposition 'the miners are in shaft A,' B stand for 'the miners are in shaft B,' b_A stand for 'shaft A is blocked,' b_B stand for 'shaft B is blocked,' and b stand for 'a shaft is blocked.' Therefore, $\odot_{Res} b \wedge \neg \odot^{S}_{Res}\, b$ holds at every index based on m_2 and m_3: *at these indices the rescuers objectively ought to have blocked some shaft, but they did not subjectively ought to have blocked some shaft.*

In my view, subjective ought-to-do's are very compelling in *kiobt*-models. The reason is that they concern states of affairs that an agent can knowingly enforce and such that the agent knows that they should be enforced (see Chapter 4's Subsection 4.5.1). Therefore, for deontic context $\odot_{\alpha}\varphi \wedge \neg \odot^{S}_{\alpha}\, \varphi$, the existence of subjectively optimal actions for which not all epistemic clusters support φ justifies the claims that—*generally speaking*—bringing about φ in this level comes with less praiseworthiness than in *Level 2* and refraining from bringing about φ in this level comes with less blameworthiness than in *Level 2*. To illustrate these claims, suppose that the miners were trapped in shaft A in the example above (Figure 6.2), and that the rescuers followed their objective sense of ought-to-do and knowingly and intentionally blocked a shaft, which only by luck turned out to be the correct one (shaft A). This means that the actual index is $\langle m_2, h_1\rangle$, where $\mathcal{M}, \langle m_2, h_1\rangle \models K_{Res}[Res]b_A \wedge I_{Res}[Res]b_A$. Regardless of the fact that the rescuers engaged in a praiseworthy mode of responsibility with respect to proposition b at $\langle m_2, h_1\rangle$, they took a risky choice (R_1) that disagrees with a subjectively optimal action (R_3) over b. Thus, the level of praiseworthiness should be lower than the one in a hypothetical case where to block a shaft were also a subjective ought-to-do. In contrast, suppose that the rescuers decided to play it safe and thus knowingly and intentionally refrained from blocking any shaft (choice R_3). One miner died because of this, and the rescuers can be excused for not saving this miner if they claim that they were not subjectively obligated to do so; they chose a subjectively optimal action that led to saving 9 miners. Therefore, the level of blameworthiness should be lower than the one in a hypothetical case where to block a shaft were also a subjective ought-to-do.[23]

The example in the previous paragraph highlights one of the problems with objective ought-to-do's that were extensively discussed in Chapter 4, where these problems revolve around the lack of a relation between an agent's objective obli-

[23] Recall that the criterion of comparison between levels, discussed in *Level 1*, is that, for a fixed conjunction of formulas characterizing a combination of sub-categories of α's responsibility with respect to φ at $\langle m, h\rangle$, the mode that results from coupling this conjunction with the deontic context of *Level x* (with x in 1–3) gets a higher degree of praiseworthiness, resp. blameworthiness, than the mode that results from coupling this conjunction with the deontic context of *Level (x+1)*.

gations and its knowledge. Notice that the rescuers objectively ought to have taken a gamble, which, given the risk of failure, is definitely not an intuitive decision to have made. The prevalence of such problems supports my view that subjective ought-to-do's are more compelling than objective ones.

Once again, and similar to what was said for the deontic context of previous levels, for $\psi \in \{\Box K_\alpha \neg \odot_\alpha \varphi, K_\alpha \neg \odot_\alpha \varphi\}$, the conjunctions $\left(\odot_\alpha \varphi \wedge \neg \odot^S_\alpha \varphi\right) \wedge \psi$ and $\left(\odot_\alpha \varphi \wedge \neg \odot^S_\alpha \varphi\right) \wedge \neg\psi$ present sub-cases of the main deontic context $\odot_\alpha \varphi \wedge \neg \odot^S_\alpha \varphi$.[24] Furthermore, for this deontic context, it might be the case that it was not even possible for α to have knowingly seen to it that φ—$\neg\Diamond K_\alpha[\alpha]\varphi$ might hold at $\langle m, h\rangle$, because $\odot_\alpha \varphi \to \Diamond K_\alpha[\alpha]\varphi$ is not valid. For situations in which $\neg\Diamond K_\alpha[\alpha]\varphi$ holds at $\langle m, h\rangle$, if α brought about φ, then it must have been unknowingly. For instance, suppose that $\neg\Diamond K_\alpha[\alpha]\varphi$ holds and that α brought about φ. Thus, α was causal-active responsible for φ at $\langle m, h\rangle$. The mode of α's responsibility with respect to φ at $\langle m, h\rangle$ is then tagged praiseworthy Low_A. However, since it was impossible for α to bring about φ knowingly, α could in principle receive a bit more praise than the one it would receive in a hypothetical case where $\Diamond K_\alpha[\alpha]\varphi$ were to also hold. Thus, for $\psi = \Diamond K_\alpha[\alpha]\varphi$, the conjunctions $\left(\odot_\alpha \varphi \wedge \neg \odot^S_\alpha \varphi\right) \wedge \psi$ and $\left(\odot_\alpha \varphi \wedge \neg \odot^S_\alpha \varphi\right) \wedge \neg\psi$ present further sub-cases of the main deontic context.

<u>Level 4</u>: when deontic context $\neg \odot_\alpha \varphi \wedge \neg \odot^S_\alpha \varphi$ holds at $\langle m, h\rangle$.

Unless α either objectively or subjectively ought have seen to it that $\neg\varphi$ at $\langle m, h\rangle$ (which would imply that a deontic context of the previous levels holds with respect to $\neg\varphi$), then in this level neither bringing about φ nor refraining from doing so elicits any interest in terms of blame-or-praise assignment. Thus, the modes of α's responsibility with respect to φ at $\langle m, h\rangle$ whose constituent (3) is deontic context $\neg \odot_\alpha \varphi \wedge \neg \odot^S_\alpha \varphi$ get a *neutral* degree of praiseworthiness/blameworthiness, provided that $\neg \odot_\alpha \neg\varphi \wedge \neg \odot^S_\alpha \neg\varphi$ also holds. This means that such modes are seen as calling for neither praise nor blame. To illustrate this intuition, think of φ as a deontically inconsequential state of affairs in the environment, such as the one mentioned in Section 6.2 (p. 268): causing a particular footprint, with no harm or foul for anyone, on a road that is busy on a daily basis. It is unlikely that a person should be blamed or praised for having caused the footprint, because both to do so and to prevent oneself from doing so were not obligations of any kind.

This concludes my discussion on the characterization of modes of responsibility in terms of *IEAUST* formulas.

[24] Observe that, when deontic context $\odot_\alpha \varphi \wedge \neg \odot^S_\alpha \varphi$ holds at $\langle m, h\rangle$, then at $\langle m, h\rangle$ α knew *ex ante* that it was not subjectively obligated to see to it that φ: formula $\neg \odot^S_\alpha \varphi \to \Box K_\alpha \neg \odot^S_\alpha \varphi$ is valid (see item 11 in the list of *EAUST*'s logic-based properties, Chapter 4, Subsection 4.5.1, p. 175).

6.4 Logic-Based Properties & Axiomatization

The logic-based properties of *IEAUST* are obtained by grouping *EAUST*'s properties (Chapter 4's Subsection 4.5.1) and *IEST*'s properties (Chapter 5's Subsection 5.4.1). Just as done in all the previous chapters, in this section I introduce proof systems. More precisely, I present two systems:

- A sound system for *IEAUST*, for which achieving a completeness result is still an open problem.
- A sound and complete system for a technical extension of *IEAUST* that I refer to as *bi-valued IEAUST*. Bi-valued *IEAUST* was devised with the aim of having a completeness result for a logic that would be reasonably similar to the one presented in Section 6.3.

As for the first bullet point, a proof system for *IEAUST* is defined as follows:

Definition 6.5 (Proof system for *IEAUST*)**.** *Let Λ_R be the proof system defined by the following axioms and rules of inference:*

- (Axioms) *All classical tautologies from propositional logic; the* **S5** *schemata for $\Box$, $[\alpha]$, and K_α; the* **KD** *schemata for I_α; and the schemata given in Table 6.6.*

Basic-stit-theory schemata: $\Box\varphi \to [\alpha]\varphi$ (*SET*) For distinct $\alpha_1, \dots, \alpha_m$, $\bigwedge_{1\leq k\leq m} \Diamond[\alpha_i]\varphi_i \to \Diamond\left(\bigwedge_{1\leq k\leq m} [\alpha_i]\varphi_i\right)$ (*IA*)	*Schemata for knowledge:* $K_\alpha\varphi \to [\alpha]\varphi$ (*OAC*) $\Diamond K_\alpha\varphi \to K_\alpha\Diamond\varphi$ (*Unif − H*)
Schemata for objective ought-to-do's: $\odot_\alpha(\varphi \to \psi) \to (\odot_\alpha\varphi \to \odot_\alpha\psi)$ (*A*1) $\Box\varphi \to \odot_\alpha\varphi$ (*A*2) $\odot_\alpha\varphi \to \Box \odot_\alpha \varphi$ (*A*3) $\odot_\alpha\varphi \to \odot_\alpha([\alpha]\varphi)$ (*A*4) $\odot_\alpha\varphi \to \Diamond[\alpha]\varphi$ (*Oic*)	*Schemata for subjective ought-to-do's:* $\odot^S_\alpha(\varphi \to \psi) \to (\odot^S_\alpha\varphi \to \odot^S_\alpha\psi)$ (*A*5) $\odot^S_\alpha\varphi \to \odot^S_\alpha(K_\alpha\varphi)$ (*A*6) $K_\alpha\Box\varphi \to \odot^S_\alpha\varphi$ (*SuN*) $\odot^S_\alpha\varphi \to \Diamond K_\alpha\varphi$ (*s.Oic*) $\odot^S_\alpha\varphi \to K_\alpha\Box \odot^S_\alpha \varphi$ (*s.Cl*) $\odot^S_\alpha\varphi \to \neg \odot_\alpha \neg\varphi$ (*ConSO*)
Schemata for intentionality: $\Box K_\alpha\varphi \to I_\alpha\varphi$ (*InN*) $I_\alpha\varphi \to \Box K_\alpha I_\alpha\varphi$ (*KI*)	

Table 6.6: *Axioms for the modalities' interactions.*

- *(Rules of inference) Modus Ponens, Substitution, and Necessitation for all modal operators.*

The reader will notice that, with the exception of (*ConSO*), all the schemata in Table 6.6 have already been extensively discussed in the previous chapters. As for (*ConSO*), it characterizes syntactically that subjective and objective ought-to-do's are consistent, something that was discussed in item 13 in the list of *EAUST*'s logic-based properties (Chapter 4, Subsection 4.5, p. 175). An important result for Λ_R is the following proposition, whose proof is relegated to Appendix E.

Proposition 6.6 (Soundness of Λ_R)**.** *The proof system Λ_R is sound with respect to the class of* kiobt-*models.*

Unfortunately, the question of whether Λ_R is complete with respect to the class of *kiobt*-models is still an open problem. Now, in the search for a complete proof system for *IEAUST*, and following a strategy found in my joint works with Jan Broersen (Abarca & Broersen, 2019, 2021a), I tried to first prove completeness of Λ_R with respect to a class of more general models, that I refer to as *bi-valued kiobt*-models (Definition 6.7 below). This strategy led to the need of dropping one of the schemata in Λ_R: (*ConSO*). More precisely, if Λ'_R is obtained from Λ_R by eliminating (*ConSO*) in Definition 6.5, then Λ'_R turns out to be sound and complete with respect to the class of *bi-valued kiobt*-models. The formal statements are included below.

Definition 6.7 (Bi-valued *kiobt*-frames & models)**.** *A tuple*

$$\left\langle M, \sqsubset, Ags, \mathbf{Choice}, \{\sim_\alpha\}_{\alpha \in Ags}, \tau, \mathbf{Value}_{\mathcal{O}}, \mathbf{Value}_{\mathcal{S}} \right\rangle$$

is called a bi-valued kiobt-*frame iff*

- *$M, \sqsubset, Ags, \mathbf{Choice}$, $\{\sim_\alpha\}_{\alpha \in Ags}$, and τ are defined just as in Definition 6.2.*
- *$\mathbf{Value}_{\mathcal{O}}$ and $\mathbf{Value}_{\mathcal{S}}$ are functions that independently assign to each history $h \in H$ a real number.*

A bi-valued kiobt-*model $\mathcal{M}$, then, results from adding a valuation function $\mathcal{V}$ to a bi-valued* kiobt-*frame, where $\mathcal{V} : P \to 2^{I(M \times H)}$ assigns to each atomic proposition of $\mathcal{L}_\mathsf{R}$ a set of indices (recall that P is the set of propositions in $\mathcal{L}_\mathsf{R}$).*

The two value functions in bi-valued *kiobt*-frames allow us to redefine the dominance orderings so that they are independent from one another, something that proves useful in achieving a completeness result in the style of Abarca and Broersen (2019). For $\alpha \in Ags$ and $m \in M$, two general orderings $\leq$ and $\leq_s$ are first defined on 2^{H_m}: for $X, Y \subseteq H_m$, $X \leq Y$, resp. $X \leq_s Y$, iff $\mathtt{Value}_{\mathcal{O}}(h) \leq \mathtt{Value}_{\mathcal{O}}(h')$, resp. $\mathtt{Value}_{\mathcal{S}}(h) \leq \mathtt{Value}_{\mathcal{S}}(h')$, for every $h \in X$ and $h' \in Y$. Then, for $\alpha \in Ags$ and

$m \in M$, an objective dominance ordering $\preceq$ is now defined on $\mathbf{Choice}^m_\alpha$ by the rule: $L \preceq L'$ iff for every $S \in \mathbf{State}^m_\alpha, L \cap S \leq L' \cap S$. In turn, for $\alpha \in Ags$ and $m \in M$, a subjective dominance ordering $\preceq_s$ is now defined on $\mathbf{Choice}^m_\alpha$ by the rule: $L \preceq_s L'$ iff for all m' such that $m \sim_\alpha m'$ and each $S \in \mathbf{State}^m_\alpha, [L]^{m'}_\alpha \cap S \leq_s [L']^{m'}_\alpha \cap S$. With these new notions, the sets $\mathbf{Optimal}^m_\alpha$ and $\mathbf{SOptimal}^m_\alpha$ are redefined accordingly, and the evaluation rules for the formulas of $\mathcal{L}_\mathsf{R}$ (with respect to bi-valued *kiobt*-models) are given just as in Definition 6.3. As mentioned before, I refer to the resulting logic as *bi-valued IEAUST*. Bi-valued *IEAUST*, then, admits the following metalogic result, whose proof is sketched in Appendix E.

Theorem 6.8 (Soundness & Completeness of Λ'_R). *Let Λ'_R be the proof system obtained from Λ_R by eliminating (ConSO) in Definition 6.5. Then Λ'_R is sound and complete with respect to the class of bi-valued* kiobt-*models.*

6.5 Conclusion

I want to conclude this chapter with a discussion of three topics. First, I discuss a possible extension of *IEAUST* with p-1 belief and with doxastic obligations. Secondly, I present a proposal for formalizing the modes of *mens rea*. Finally, I advance some comments on group-related notions and on collective responsibility.

6.5.1 Extension with P-1 Belief & Doxastic Obligations

As mentioned frequently in this thesis, belief is an important epistemic component of responsibility. In Chapter 4's conclusion (Subsection 4.6.1), I extended epistemic act-utilitarian stit theory (*EAUST*) with p-1 belief and with doxastic ought-to-do's. The idea was that agent α doxastically ought to have brought about φ iff φ is an effect of α's choices of action that maximize expected deontic utility (see Definitions 4.32 and 4.33 on p. 183). It is clear that the doxastic dimension—with the sense of ought-to-do that corresponds to it—only adds interesting nuances to systematic blame-or-praise assignment.

For instance, suppose that in the *Miners Paradox* the rescuers thought that they heard voices coming out of shaft A. In consequence, the probability assigned to the miners' being trapped in shaft A was much higher than the one assigned to their being trapped in shaft B. Suppose further that, as a result of this, the choice of blocking shaft A was the only one maximizing the rescuers' expected deontic utility. Thus, according to the semantics for doxastic ought-to-do's on p. 184, the rescuers were doxastically obligated to block shaft A. If b_A stands for the proposition 'shaft A is blocked,' then formula $\odot^{\mathcal{B}}_{Res} b_A$ holds. Suppose, then, that the rescuers indeed block shaft A, and that this was a fatal mistake due to the fact that the miners were actually trapped in shaft B. In this case, although 10 miners died, one could say that the degree of blameworthiness of the rescuers is less than the one in a hypothetical case where to block a shaft were not a doxastic ought-to-do (for which formula $\neg \odot^{\mathcal{B}}_{Res} b_A$ would hold).

Furthermore, an important variation of the category of informational responsibility results from using a belief modality $B_\alpha\varphi$ in place of $K_\alpha\varphi$. Certainly, there are situations where one would claim that agent α was informational-active responsible for φ iff, for instance, $[\alpha]\varphi \wedge B_\alpha[\alpha]\varphi \wedge B_\alpha\Diamond[\alpha]\neg\varphi$ holds, because this formula means that *α saw to it that φ, α believed that it saw to it that φ, and α believed that it was possible to prevent φ*. Similarly, one could very well claim that α was informational-passive responsible for φ iff $\varphi \wedge B_\alpha\neg[\alpha]\neg\varphi \wedge B_\alpha\Diamond[\alpha]\neg\varphi$ holds, because this formula means that *α refrained from preventing φ, α believed that it refrained from preventing φ, and α believed that it was possible to prevent φ.*

Therefore, extending *IEAUST* with the p-1 belief modality $B_\alpha\varphi$ and with $\odot^{\mathcal{B}}_\alpha\varphi$ leads to a framework that accommodates a wide variety of new modes of responsibility. To be precise, let $\mathcal{L}_{\mathsf{R}}{}'$ be obtained by extending $\mathcal{L}_{\mathsf{R}}$ with the aforementioned doxastic modalities, and let $\mathcal{M}$ be a finite *kiobt*-model to which a set $\{\mu_\alpha\}_{\alpha\in Ags}$ of probability functions is added. For $\alpha \in Ags$, let μ_α underlie the semantics for $B_\alpha\varphi$ according to the semantics for p-1 belief on p. 112, and let the semantics for $\odot^{\mathcal{B}}_\alpha\varphi$ be given just as on p. 184. Then, for every φ of $\mathcal{L}'_{\mathsf{R}}$, $\alpha \in Ags$, there are now 8 main possible deontic contexts, according to whether $\Delta\varphi$ or $\neg\Delta\varphi$ holds at $\langle m, h\rangle$, where $\Delta \in \left\{\odot_\alpha, \odot^{\mathcal{S}}_\alpha, \odot^{\mathcal{B}}_\alpha\right\}$. Furthermore, informational responsibility now includes a new sub-category, where B_α is used instead of K_α in Table 6.1. To illustrate this extension, let me formalize the aforementioned example based on the *Miners Paradox*, where the rescuers assigned a much higher probability to the miners' being trapped in shaft A than the one assigned to their being trapped in shaft B. Consider Figure 6.3 below, which expands Figure 6.2.

Since the rescuers heavily lean toward believing that the miners are trapped in shaft A, let μ_{Res} be such that (a) $\mu_{Res}\left(\{\langle m_2, h\rangle\} \mid \pi_{Res}\left[\langle m_2, h\rangle\right]\right) = .91$ for every $h \in H_{m_2}$, and (b) $\mu_{Res}\left(\{\langle m_3, h\rangle\} \mid \pi_{Res}\left[\langle m_3, h\rangle\right]\right) = .09$ for every $h \in H_{m_3}$. As such, $EU^{\langle m_2,h\rangle}_{Res}(R_1) = 9.1 = EU^{\langle m_3,h'\rangle}_{Res}(R_4)$, $EU^{\langle m_2,h\rangle}_{Res}(R_2) = .9 = EU^{\langle m_3,h'\rangle}_{Res}(R_5)$, and $EU^{\langle m_2,h\rangle}_{Res}(R_3) = 9 = EU^{\langle m_3,h'\rangle}_{Res}(R_6)$ for every $h \in H_{m_2}$ and $h' \in H_{m_3}$. This implies that $\mathbf{EU}^{\langle m_2,h\rangle}_{Res} = \{R_1\}$ for every $h \in H_{m_2}$, and that $\mathbf{EU}^{\langle m_3,h'\rangle}_{Res} = \{R_4\}$ for every $h' \in H_{m_3}$. As for the other deontic modalities, observe that $\mathbf{Optimal}^{m_2}_{Res} = \{R_1\}$, that $\mathbf{Optimal}^{m_3}_{Res} = \{R_5\}$, that $\mathbf{SOptimal}^{m_2}_{Res} = \{R_1, R_2, R_3\}$, and that $\mathbf{SOptimal}^{m_3}_{Res} = \{R_4, R_5, R_6\}$.

Just as in Figure 6.2, in Figure 6.3 b_A stands for 'shaft A is blocked,' b_B stands for 'shaft B is blocked,' and b stands for 'a shaft is blocked.' Therefore, formulas $\odot_{Res}b$,

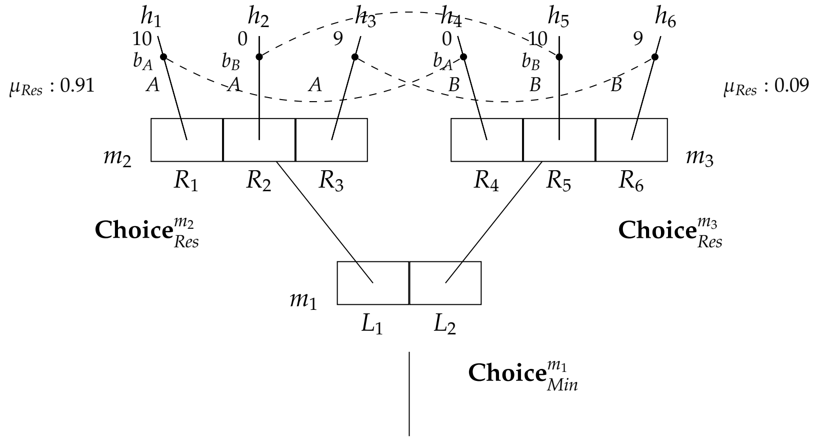


Figure 6.3: *Miners paradox with doxastic notions.*

$\neg\odot^{\mathcal{S}}_{Res} b$, $\odot^{\mathcal{B}}_{Res} b_A$, and $\odot^{\mathcal{B}}_{Res} b$ hold at every index based on m_2 and m_3: *at these indices the rescuers objectively ought to have blocked some shaft, doxastically ought to have blocked shaft A, and did not subjectively ought to have blocked some shaft.* Since blocking shaft A implies blocking some shaft ($\mathcal{M} \models b_A \to b$), then $\mathcal{M}, \langle m_i, h\rangle \models \odot_{Res} b \wedge \neg\odot^{\mathcal{S}}_{Res} b \wedge \odot^{\mathcal{B}}_{Res} b$ for all $i \in \{2, 3\}$ and $h \in H_{m_i}$: *at all indices based on m_2 and m_3 the rescuers objectively and doxastically ought to have blocked a shaft, but they were not subjectively obligated to do so.*

Thus, this is an example where the deontic context is of the form $\odot_\alpha \varphi \wedge \neg \odot^{\mathcal{S}}_{\alpha} \varphi \wedge \odot^{\mathcal{B}}_{\alpha} \varphi$.[25] The modes of responsibility associated with such a context, then, could in principle be obtained by extending Tables 6.2 and 6.3, resp. Tables 6.4 and 6.5, with modes for doxastic-informational-active responsibility (characterized with $[\alpha]\varphi \wedge B_\alpha[\alpha]\varphi \wedge B_\alpha\Diamond[\alpha]\neg\varphi$), resp. with modes for doxastic-informational-passive responsibility (characterized with $\varphi \wedge B_\alpha\neg[\alpha]\neg\varphi \wedge B_\alpha\Diamond[\alpha]\neg\varphi$).

Although the extension of *IEAUST* with p-1 belief and with doxastic ought-to-do's has already been presented in my joint work with Jan Broersen (Abarca & Broersen, 2022), a full-fledged exploration of the logic in the context of responsibility attribution—as well as an exploration of its metalogic results—remains to be done.

[25] Recall from Footnote 31 (p. 186) that subjective ought-to-do's are not necessarily consistent with doxastic ones, something that highlights the discrepancy between the principle of subjective dominance and the principle of maximization of expected deontic utility. Thus, the incorporation of p-1 belief and doxastic ought-to-do's into *IEAUST* opens up many possibilities for a complex analysis of deontic contexts and of their associated levels of praiseworthiness/blameworthiness.

6.5.2 Formalizing the Modes of *Mens Rea*

Recall that Broersen's (2011a) initial motivation for categorizing the notion of responsibility was *mens rea*. As it turns out, the ideas behind Subsection 6.3.3's stit-theoretic formalization of modes of responsibility can be used to also formalize the modes of *mens rea* that were introduced on p. 266.

Suppose that φ is of the form $\neg\psi$ for a formula ψ that stands for an illegal outcome, or a criminal offense, at index $\langle m,h\rangle$. Thus, the deontic context that holds at $\langle m,h\rangle$ will most likely be included in one of *Levels* 1–3. In other words, $\mathcal{M},\langle m,h\rangle \models \odot_\alpha\varphi \vee \odot^S_\alpha\varphi$. For any of the implied deontic contexts, one can characterize the *mens rea* mode *purposefully*, for criminal agent α, with formula $\Box K_\alpha\left(\odot_\alpha\varphi \vee \odot^S_\alpha\varphi\right) \wedge (K_\alpha[\alpha]\neg\varphi \wedge I_\alpha[\alpha]\neg\varphi)$. This formula holds at $\langle m,h\rangle$ iff at this index (a) α knew *ex ante* that to see to it that $\neg\varphi$ was prohibited on some deontic account, and (b) provided that $\Diamond K_\alpha[\alpha]\varphi$ also holds at the index, α was causal-active, informational-active, and motivational-active responsible for $\neg\varphi$ at $\langle m,h\rangle$. Thus, the *mens rea* mode *purposefully* is a mode of α's responsibility that gets a high degree of blameworthiness (with respect to the deontic context at hand).

Similarly, one can characterize the *mens rea* mode *knowingly* with formula $\Box K_\alpha\left(\odot_\alpha\varphi \vee \odot^S_\alpha\varphi\right) \wedge K_\alpha[\alpha]\neg\varphi$. This formula holds at $\langle m,h\rangle$ iff at this index (a) α knew *ex ante* that to see to it that $\neg\varphi$ was prohibited on some deontic account, and (b) provided that $\Diamond K_\alpha[\alpha]\varphi \wedge \neg I_\alpha\neg[\alpha]\varphi$ also holds at the index, α was causal-active, informational-active, and not motivational-passive for $\neg\varphi$ at the index. Thus, this version of the *mens rea* mode *knowingly*—where there is a complete lack of intent—is a mode of α's responsibility that gets a middle degree of blameworthiness (with respect to the deontic context at hand).

As for the *mens rea* mode *recklessly*, one can use an extension of *IEAUST* with belief, just as the one presented in the previous subsection, to formalize it. To clarify, formula $\Box B_\alpha\left(\odot_\alpha\varphi \vee \odot^S_\alpha\varphi\right) \wedge [\alpha]\theta \wedge \Box([\alpha]\theta \rightarrow \neg\varphi) \wedge B_\alpha\Diamond([\alpha]\theta \rightarrow \neg\varphi)$ is a good candidate for characterizing the mode *recklessly*. This formula holds at $\langle m,h\rangle$ iff at this index (a) α believed—regardless of anyone's choice—that to see to it that $\neg\varphi$ was prohibited on some deontic account, (b) α causally brought about θ such that it was settled that $\neg\varphi$ follows from α's seeing to it that θ, and (c) α believed that it was possible that its bringing about θ could have implied $\neg\varphi$. In this case, α was causal-active responsible for θ at $\langle m,h\rangle$. The validity of schema (*SET*) and of schema (*K*) for $[\alpha]$ yields that α was also causal-active responsible for $\neg\varphi$ at $\langle m,h\rangle$. Provided that $\Diamond K_\alpha[\alpha]\varphi \wedge \neg I_\alpha\neg[\alpha]\varphi \wedge \neg K_\alpha\neg[\alpha]\varphi$ also holds at the index, α was neither informational-passive nor motivational-passive for $\neg\varphi$.

Thus, this version of the *mens rea* mode *recklessly*—where there is a complete lack of intent and of knowledge—is a mode of α's responsibility that gets a low degree of blameworthiness (with respect to the deontic context at hand).

As for the *mens rea* mode *negligently*, formula $\Box K_\beta \left(\odot_\alpha\varphi \vee \odot^S_\alpha\varphi\right) \wedge [\alpha]\theta \wedge \Box K_\beta \left([\alpha]\theta \rightarrow \neg\varphi\right)$, where β represents a 'reasonable agent,' is a good candidate for characterizing it. This formula holds at $\langle m, h\rangle$ iff at this index (a) a reasonable agent β would have known *ex ante* that $\neg\varphi$ was prohibited on some deontic account, (b) α causally brought about θ such that $\neg\varphi$ follows from α's seeing to it that θ, and (c) a reasonable agent β would have known *ex ante* about such an implication. Just as for the mode *recklessly*, here α was causal-active responsible for θ and for $\neg\varphi$ at $\langle m, h\rangle$. Provided that $\Diamond K_\alpha[\alpha]\varphi \wedge \neg I_\alpha\neg[\alpha]\varphi \wedge \neg K_\alpha\neg[\alpha]\varphi$ also holds at the index, α was neither informational-passive nor motivational-passive for $\neg\varphi$. Thus, this version of the *mens rea* mode *negligently*—where again there is a complete lack of intent and of knowledge—also gets a low degree of blameworthiness (with respect to the deontic context at hand).

Strict liability offenses are charged and tried without appealing to any *mens rea* mental state. Typically, offenses of this kind are divided in two main categories (see, for instance, Green, 2005; Larkin JR, 2014): (1) minor infractions (such as speeding, overtime parking, or not signaling for a turn), for which the justification of reaching verdicts without requiring any proof of *mens rea* is made on the grounds of regulatory expediency; and (2) serious crimes that pose a danger to society (such as statutory rape or felony murder), for which conviction without any proof of *mens rea* is justified on the grounds of maximizing the deterrent effect of the penalty. For both categories, and if φ is of the form $\neg\psi$ for a strict liability offense ψ, one can characterize the mode *strict liability*—for criminal agent α—using α's causal-active responsibility for $\neg\varphi$ (ψ). In other words, against a deontic context that implies that $\odot_\alpha\varphi \vee \odot^S_\alpha\varphi$ holds, α's strict liability for having seen to it that $\neg\varphi$ can be characterized with formula $[\alpha]\neg\varphi \wedge \Diamond[\alpha]\varphi$.

6.5.3 Collective Responsibility

As mentioned in Subsection 4.6.2 of Chapter 4's conclusion, *collective responsibility* refers to a relation between a group of agents and some state of affairs such that the group is responsible for the state of affairs iff the group's degree of involvement in the realization of that state warrants collective blame or collective praise.

Just as in the case of individual responsibility, one can both decompose and classify collective responsibility. On the one hand, the list of components would include group agency, group knowledge & belief, group intentions & plans, and group obligations. On the other, the categories would once again be causal, informational, and motivational responsibility. Coupled with different senses of group obligations (objective, subjective, doxastic, etc.), these categories would lead to diverse modes of collective responsibility, in a similar fashion to Section 6.3's.

Adapting Section 6.3's exploration to the case of groups is a feasible endeavor, because the literature includes many paradigms to formalize group agency (Broersen, 2011a; Broersen et al., 2006b; Herzig & Schwarzentruber, 2008; Lorini, 2013; Lorini et al., 2014; Payette, 2014; Schwarzentruber, 2012; Tamminga, 2013) (see also Chapter 2's Subsection 2.4.1), group knowledge & belief (Barwise, 1989; Fagin et al., 1995; Gerbrandy, 1998; Halpern & Fagin, 1989), group intentions & plans (Duijf, 2018, Chapters 2& 3;Bratman, 2013), and group obligations (Horty, 2001, Chapter 6;Tamminga, 2013) (see also Chapter 4's Subsection 4.6.2).

For instance, take $G \subseteq Ags$. Let group agency be defined exactly as in Chapter 2's Subsection 2.4.1, so that modality $[G]\varphi$ is based on $\mathbf{Choice}^m_G := \{\bigcap_{\alpha\in G}\mathbf{Choice}^m_\alpha(h); h \in H_m\}$. Let group knowledge be distributed, so that modality $D_G\varphi$ is based on $\bigcap_{\alpha\in G}\sim_\alpha$. Let group p-d intentions be obtained from individual p-d intentions according to the rules of the joint topology, so that modality $I_G\varphi$ is based on the topology generated by $\bigcup_{\alpha\in G}\tau_\alpha^{\langle m,h\rangle}$. Finally, let group objective and subjective obligations be defined just as in Subsection 4.6.2 of Chapter 4's conclusion, so that modality $\odot_G\varphi$, resp. $\odot^S_G\varphi$, is based on an objective, resp. subjective, dominance ordering of joint actions. Then, first of all, one can advance the following syntactic characterizations: formula $[G]\varphi \wedge \Diamond[G]\neg\varphi$ for collective causal-active responsibility (with respect to φ), and $\varphi \wedge \neg[G]\neg\varphi \wedge \Diamond[G]\neg\varphi$ for collective causal-passive responsibility; formula $D_G[G]\varphi \wedge \Diamond D_G[G]\neg\varphi$ for collective informational-active responsibility, and $\varphi \wedge D_G\neg[G]\neg\varphi \wedge \Diamond D_G[G]\neg\varphi$ for collective informational-passive responsibility; formula $[G]\varphi \wedge I_G[G]\varphi \wedge \Diamond D_G[G]\neg\varphi$ for collective motivational-active responsibility, and $\varphi \wedge I_G\neg[G]\neg\varphi \wedge \Diamond D_G[G]\neg\varphi$ for collective motivational-passive responsibility. Secondly, coupling sub-categories of these forms with the deontic contexts given by whether $\Delta\varphi$ or $\neg\Delta\varphi$ is satisfied at a given index, where $\Delta \in \left\{\odot_G, \odot^S_G\right\}$, one can formalize levels of collective praiseworthiness/blameworthiness with analogs of Tables 6.2 and 6.3.

Now, if extending *IEAUST* with group notions for the formalization of collective responsibility is feasible, the big challenge comes from choosing those group notions' semantics so that one can successfully—and systematically—assess the relations between individual and collective responsibility. What principles will be validated? What principles *does one wish* to be validated?

To illustrate this challenge, consider the question of whether *responsibility voids* exist. Think of the following version of the *discursive dilemma*, atrributed to Duijf (2018, Introduction): suppose that a committee of academics, consisting of *Marie*, *Mel*, and *Mo*, is deciding on whether to award tenure to *Mr.Borderline*. The university's tenure policy requires excellence in research, service, and teaching. Thus, the committee is to decide on awarding tenure by first voting on each of these fields of competence, then aggregating its members' votes by majority, and finally deriving the collective decision in line with university rules. Suppose that the members vote in accordance with Table 6.7.

	Research *r*	Service *s*	Teaching *t*	Tenure? *r* & *s* & *t*
Marie	Yes	Yes	No	No
Mel	No	Yes	Yes	No
Mo	Yes	No	Yes	No
Group	Yes	Yes	Yes	→Yes / ↓No

Table 6.7: *The discursive dilemma.*

As such, the committee collectively decided to award tenure even when they were unanimously opposed to doing so. Now, suppose further that *Mr. Borderline* turns out to be a terrible choice for the university. Is the committee collectively responsible/blameworthy for awarding tenure? Are any of the committee members individually responsible/blameworthy for contributing to awarding tenure? Is this a case where a group is collectively blameworthy for an outcome without any of its members being individually blameworthy for it? Or, in other words, is there a responsibility void here? Well, it all depends on the characterizations of individual and collective responsibility that one wishes to adopt.

For instance, suppose that group agency is obtained by aggregating decisions according to the majority rule. Then one can say that the committee is collectively causally responsible for *Mr. Borderline*'s tenure, although neither member is individually causally responsible for such a decision. However, what about informational responsibility? The answer to whether the committee and/or its members are informationally responsible for tenure depends on the kind of group knowledge considered. For instance, if knowledge is aggregated by the rules of distributed knowledge (Fagin et al., 1995; Halpern & Fagin, 1989), and if communication is possible between the members of the committee, then one could say that the group is collectively informationally responsible, and that its members are individually informationally responsible as well. If communication were not possible, then one would say that neither the group nor its members are informationally responsible—unless one of the members somehow knew how the others would vote (in which case that member is informationally responsible, the group is implicitly informationally responsible, and the group is not explicitly informationally responsible (Gerbrandy, 1998)).

A similar reasoning applies to motivational responsibility. The answer to whether the committee and/or its members are motivationally responsible for tenure depends on the kind of collective intentionality considered. Suppose that the committee members agree that their collective intention is to award tenure. If the individual intentions are *team-directed* (see Duijf, 2018, Chapter 5), meaning that each member intends to play their part so that the collective intention is achieved, then both the group and its members are motivationally responsible. If there is agreement that the collective intention is to award tenure but a member has a non-team-directed intention that tenure is not awarded, then the group is collectively motivationally responsible, but said member is not. Suppose, in contrast, that the group's plan or intention is obtained by aggregating the individual intentions that are implied by the votes in Table 6.7 (which are not cooperative or team-directed), then neither the group is collectively motivationally responsible nor any of its members is individually motivationally responsible for awarding tenure.

As the reader can foresee, the individual-collective relation is immensely complex, and there are quite a few paths to choose from in order to formalize it. This is a very interesting line for future work, and the reader is referred to Duijf (2018, 2022) for germane reviews, proposals, and discussions.

Appendix E Metalogic Results for *IEAUST*

E.1 Soundness

Proposition E.1 (Soundness of Λ_R)**.** *The system* Λ_R *(Definition 6.4) is sound with respect to the class of* kiobt-*models.*

Proof. The proof of soundness is routine: the validity of the **S5** schemata for $\Box$ and $[\alpha]$, as well as that of (*SET*) and (*IA*), is standard from *BST*; the validity of the **S5** schemata for K_α is standard from *EST*; the validity of (*OAC*) and ($Unif - H$) is shown exactly as in Chapter 4's Proposition C.40; the validity of schemata ($A1$)–($A4$), as well as that of (*Oic*), is standard from *AUST* (Murakami, 2004), and can be shown just as in Chapter 4 (Theorem 4.28 and Proposition C.38); the validity of schemata ($A5$) and ($A6$), as well as that of (*SuN*), (*s.Oic*), (*s.Cl*), and (*ConSO*) can be shown just as in Chapter 4 (Proposition C.40 and Observation C.39); the validity of the **KD** schemata for I_α, as well as that of (*InN*), follows from Definitions 6.2, 5.4, and 5.5; and the validity of (*KI*) follows from frame condition ($\mathtt{KI}$). □

E.2 Completeness

As mentioned in the main body of the chapter, whether Λ_R is complete with respect to the class of *kiobt*-models is still an open problem. However, the proof system Λ'_R—obtained from Λ_R by eliminating (*ConSO*) in Definition 6.5—is sound and complete with respect to the class of bi-valued *kiobt*-models (Definition 6.7). Soundness follows from Proposition E.1, and the proof of completeness is obtained by integrating the proofs of completeness in Chapters 4 and 5. More precisely, the proof of completeness will be sketched below as a two-step process. First, I introduce a Kripke semantics for *bi-valued IEAUST*, where the formulas of $\mathcal{L}_\mathsf{R}$ are evaluated on bi-valued Kripke-*kios*-models (Definition E.2). Completeness of $\Lambda_R{}'$ with respect to the class of these structures is shown via the well-known technique of canonical models. Secondly, a truth-preserving correspondence between bi-valued Kripke-*kios*-models and a sub-class of bi-valued *kiobt*-models is used to prove completeness with respect to bi-valued *kiobt*-models via completeness with respect to bi-valued Kripke-*kios*-models.

A Kripke semantics for *IEAUST* is defined as follows:

Definition E.2 (Bi-valued Kripke-*kios*-frames & models)**.** *A tuple*

$$\left\langle W, Ags, R_\Box, \mathbf{Choice}, \{\approx_\alpha\}_{\alpha\in Ags}, \left\{R^I_\alpha\right\}_{\alpha\in Ags}, \mathtt{Value}_\mathcal{O}, \mathtt{Value}_\mathcal{S}\right\rangle$$

is called a bi-valued *Kripke-*kios*-frame iff*

- *$W, Ags, R_\Box, Ags,$ $\mathtt{Choice}$, $\{\approx_\alpha\}_{\alpha\in Ags}$, and $\left\{R^I_\alpha\right\}_{\alpha\in Ags}$ are defined just as in Chapter 5's Definition D.14 (p. 253), which implies that the non-intentional attributes are defined just as in Chapter 4's Definition C.41 (p. 200).*

  Thus, recall from Definition C.41 that, for $\beta \in Ags$ and $w \in W$, $\mathtt{State}^{\overline{w}}_\beta := \left\{S \subseteq \overline{w}; S = \bigcap_{\alpha\in Ags-\{\beta\}} s(\alpha), \text{for } s \in \mathtt{Select}^{\overline{w}}\right\}$, where $\mathtt{Select}^{\overline{w}}$ denotes the set of all selection functions at $\overline{w}$ (i.e., functions that assign to each α a member of $\mathtt{Choice}^{\overline{w}}_\alpha$). Also from Definition C.41 recall that, for $w, v \in W$ such that $w \approx_\alpha v$ and $L \in \mathtt{Choice}^{\overline{w}}_\alpha$, L's epistemic cluster at $\overline{v}$ is the set $[\![L]\!]^{\overline{v}}_\alpha := \{u \in \overline{v};\ \textit{there is } o \in L \text{ such that } o \approx_\alpha u\}$.

  Recall from Definition D.14 (p. 253) that, for $\alpha \in Ags$ and $w \in W$, α's ex ante information set *at w is defined as $\pi^\Box_\alpha[w] := \{v; w \approx_\alpha \circ R_\Box v\}$. Also from Definition D.14 recall that, for $x \in X$, $x \uparrow_{R^{I+}_\alpha}$ denotes the set $\left\{y \in X \mid xR^{I+}_\alpha y\right\}$, where R^{I+}_α denotes the reflexive closure of R^{I+}_α.*

- *$\mathtt{Value}_{\mathcal{O}}$ and $\mathtt{Value}_{\mathcal{S}}$ are functions that independently assign to each world $w \in W$ a real number.*

  These functions are used to define an objective ordering $\preceq$ and a subjective ordering $\preceq_s$ of choices. Formally, for $\alpha \in Ags$ and $w \in W$, one first defines two general orderings $\leq$ and $\leq_s$ on 2^W by the rules: $X \leq Y$ iff $\mathtt{Value}_{\mathcal{O}}(w) \leq \mathtt{Value}_{\mathcal{O}}(w')$ for all $w \in X$ and $w' \in Y$; and $X \leq_s Y$ iff $\mathtt{Value}_{\mathcal{S}}(w) \leq \mathtt{Value}_{\mathcal{S}}(w')$ for all $w \in X$ and $w' \in Y$. An objective dominance ordering $\preceq$ is then defined on $\mathtt{Choice}^{\overline{w}}_\alpha$ by the rule: $L \preceq L'$ iff $L \cap S \leq L' \cap S$ for every $S \in \mathtt{State}^{\overline{w}}_\alpha$. In turn, a subjective dominance ordering $\preceq_s$ is then defined on $\mathtt{Choice}^{\overline{w}}_\alpha$ by the rule: $L \preceq_s L'$ iff $[\![L]\!]^{\overline{v}}_\alpha \cap S \leq_s [\![L']\!]^{\overline{v}}_\alpha \cap S$ for every v such that $w \approx_\alpha v$ and every $S \in \mathtt{State}^{\overline{v}}_\alpha$. I write $L \prec L'$ iff $L \preceq L'$ and $L' \not\preceq L$, and I write $L \prec_s L'$ iff $L \preceq_s L'$ and $L' \not\preceq_s L$, so that $\mathtt{Optimal}^{\overline{w}}_\alpha := \left\{L \in \mathtt{Choice}^{\overline{w}}_\alpha; \text{ there is no } L' \in \mathtt{Choice}^{\overline{w}}_\alpha \text{ s. t. } L \prec L'\right\}$ and $\mathtt{SOptimal}^{\overline{w}}_\alpha := \left\{L \in \mathtt{Choice}^{\overline{w}}_\alpha; \text{ there is no } L' \in \mathtt{Choice}^{\overline{w}}_\alpha \text{ s. t. } L \prec_s L'\right\}$.

*A Kripke-*kios*-model $\mathcal{M}$ consists of the tuple that results from adding a valuation function $\mathcal{V}$ to a Kripke-*kios*-frame, where $\mathcal{V} : P \to 2^W$ assigns to each atomic proposition a set of worlds (recall that P is the set of propositions in $\mathcal{L}_{\mathsf{R}}$).*

Kripke-*kios*-models allow us to evaluate the formulas of $\mathcal{L}_{\mathsf{R}}$ with semantics that are analogous to the ones provided for *kiobt*-models:

Definition E.3 (Evaluation rules on Kripke models)**.** *Let $\mathcal{M}$ be a Kripke-*kios*-model, the semantics on $\mathcal{M}$ for the formulas of $\mathcal{L}_{\mathsf{KO}}$ are defined recursively by the following truth* *conditions, evaluated at world w:*

$$\begin{array}{lll}
\mathcal{M}, w \models p & \textit{iff} & w \in \mathcal{V}(p) \\
\mathcal{M}, w \models \neg\varphi & \textit{iff} & \mathcal{M}, w \not\models \varphi \\
\mathcal{M}, w \models \varphi \wedge \psi & \textit{iff} & \mathcal{M}, w \models \varphi \textit{ and } \mathcal{M}, w \models \psi \\
\mathcal{M}, w \models \Box\varphi & \textit{iff} & \textit{for each } v \in \overline{w}, \mathcal{M}, v \models \varphi \\
\mathcal{M}, w \models [\alpha]\varphi & \textit{iff} & \textit{for each } v \in \mathtt{Choice}^{\overline{w}}_{\alpha}(w), \mathcal{M}, v \models \varphi \\
\mathcal{M}, w \models K_{\alpha}\varphi & \textit{iff} & \textit{for each } v \textit{ s. t. } w \approx_{\alpha} v, \mathcal{M}, v \models \varphi \\
\mathcal{M}, w \models I_{\alpha}\varphi & \textit{iff} & \textit{there exists } x \in \pi^{\Box}_{\alpha}[w] \textit{ s. t. } x \uparrow_{R^{I+}_{\alpha}} \subseteq |\varphi| \\
\mathcal{M}, w \models \odot_{\alpha}\varphi & \textit{iff} & \textit{for all } L \in \mathtt{Choice}^{\overline{w}}_{\alpha} \textit{ s. t. } \mathcal{M}, v \not\models \varphi \textit{ for some } v \in L, \textit{ there is} \\
 & & L' \in \mathtt{Choice}^{\overline{w}}_{\alpha} \textit{ s. t. } L \prec L' \textit{ and, if } L'' = L' \textit{ or } L' \preceq_s L'', \\
 & & \textit{then } \mathcal{M}, w' \models \varphi \textit{ for every } w' \in L''_{\alpha} \\
\mathcal{M}, w \models \odot^{\mathcal{S}}_{\alpha}\varphi & \textit{iff} & \textit{for all } L \in \mathtt{Choice}^{\overline{w}}_{\alpha} \textit{ s. t. } \mathcal{M}, v \not\models \varphi \textit{ for some } w' \textit{ s. t. } w \approx_{\alpha} w' \\
 & & \textit{and some } v \in [\![L]\!]^{w'}_{\alpha}, \textit{ there is } L' \in \mathtt{Choice}^{\overline{w}}_{\alpha} \textit{ s. t. } L \prec_s L' \\
 & & \textit{and, if } L'' = L' \textit{ or } L' \preceq_s L'', \textit{then } \mathcal{M}, w''' \models \varphi \textit{ for every } w'' \\
 & & \textit{s. t. } \overline{w} \approx_{\alpha} \overline{w''} \textit{ and every } w''' \in [\![L'']\!]^{w''}_{\alpha},
\end{array}$$

where I write $|\varphi|$ to refer to the set $\{w \in W; \mathcal{M}, w \models \varphi\}$. Satisfiability, validity, and general validity are defined as usual.

A truth-preserving correspondence between Kripke-*kios*-models and *kiobt*-models is shown as follows:

Definition E.4 (Associated *kiobt*-frame)**.** *Let*

$$\mathcal{F} = \left\langle W, Ags, R_{\Box}, \mathtt{Choice}, \{\approx_{\alpha}\}_{\alpha \in Ags}, \left\{R^{I}_{\alpha}\right\}_{\alpha \in Ags}, \mathtt{Value}_{\mathcal{O}}, \mathtt{Value}_{\mathcal{S}} \right\rangle$$

*be a bi-valued Kripke-*kios*-frame.*

Then $\mathcal{F}^{T} := \left\langle M_W, \sqsubset, Ags, \mathbf{Choice}, \{\sim_{\alpha}\}_{\alpha \in Ags}, \tau, \mathbf{Value}_{\mathcal{O}}, \mathbf{Value}_{\mathcal{S}} \right\rangle$ is called the bi-valued kiobt*-frame associated with $\mathcal{F}$ iff*

- *$M_W, \sqsubset,$ **Choice**, $\{\sim_{\alpha}\}_{\alpha \in Ags}$, and τ are defined just as in Chapter 5's Definition D.19 (p. 255).*

- ***Value**$_{\mathcal{O}}$ and **Value**$_{\mathcal{S}}$ are defined by the following rules: for $h_v \in H$, $\mathbf{Value}_{\mathcal{O}}(h_v) = \mathtt{Value}_{\mathcal{O}}(v)$, and $\mathbf{Value}_{\mathcal{S}}(h_v) = \mathtt{Value}_{\mathcal{S}}(v)$.*

Proposition E.5. *Let $\mathcal{F}$ be a bi-valued Kripke-*kios*-frame. Then $\mathcal{F}^{T}$ is a bi-valued* kiobt*-frame, indeed.*

Proof. Follows from Chapter 4's Proposition C.44 (p. 202), Chapter 5's Proposition D.20 (p. 255), and Definition E.4. □

Lemma E.6. *Let $\mathcal{M}$ be a bi-valued Kripke-*kios*-model, and let $\mathcal{M}^T$ be its associated bi-valued* kiobt*-model. For all $\alpha \in Ags$, $w \in W$, and $L, N \in \texttt{Choice}^{\overline{w}}_{\alpha}$, the following conditions hold:*

(a) $L \leq N$ *iff* $L^T \leq N^T$ *and* $L \prec N$ *iff* $L^T \prec N^T$.

(b) $L \leq_s N$ *iff* $L^T \leq_s N^T$ *and* $L \prec_s N$ *iff* $L^T \prec_s N^T$.

(c) $L \in \texttt{Optimal}^{\overline{w}}_{\alpha}$ *iff* $L^T \in \mathbf{Optimal}^{\overline{w}}_{\alpha}$.

(d) $L \in \texttt{S} - \texttt{Optimal}^{\overline{w}}_{\alpha}$ *iff* $L^T \in \mathbf{S} - \mathbf{Optimal}^{\overline{w}}_{\alpha}$.

Proof. For the proofs of items b and d, see Chapter 4's Lemma C.45 (p. 203). The proofs of a and c are analogous, and the reader is referred to the proof of Proposition 4 in `https://doi.org/10.48550/arXiv.1903.10577` for details (see also Abarca & Broersen, 2019). □

Proposition E.7 (Truth-preserving correspondence). *Let $\mathcal{M}$ be a bi-valued Kripke-*kios*-model, and let $\mathcal{M}^T$ be its associated bi-valued* kiobt*-model. For all φ of $\mathcal{L}_{\mathsf{R}}$ and $w \in W$, $\mathcal{M}, w \models \varphi$ iff $\mathcal{M}^T, \langle \overline{w}, h_w \rangle \models \varphi$.*

Proof. We proceed by induction on the complexity of φ. For the base case, the cases of Boolean connectives, and the cases of all modal operators except I_α and except $\odot_\alpha$, the proofs are exactly the same as their analogs' in Chapter 4's Proposition C.46 (p. 204), using Lemma E.6 b in the case of $\odot^{S}_{\alpha}$. For the case of I_α, the proof is the same as its analog in Chapter 5's Proposition D.21 (p. 256). As for the case of $\odot_\alpha$, it follows from Lemma E.6 a according to Proposition 4 in `https://doi.org/10.48550/arXiv.1903.10577`.

□

Thus, completeness with respect to bi-valued *kiobt*-models is proved with Propositions E.8 and E.9 below.

Proposition E.8 (Completeness w.r.t. bi-valued Kripke-*kios*-models). *The proof system $\Lambda_R{}'$ is complete with respect to the class of bi-valued Kripke-*kios*-models.*

Proof. Completeness with respect to bi-valued Kripke-*kios*-models is shown via canonical models. To be precise, one defines a structure $\mathcal{M} := \left\langle W^{\Lambda'_R}, R_\Box, \texttt{Choice}, \{\approx_\alpha\}_{\alpha \in Ags}, \left\{R^I_\alpha\right\}_{\alpha \in Ags} \texttt{Value}_{\mathcal{O}}, \texttt{Value}_{\mathcal{S}}, \mathcal{V} \right\rangle$, where $W^{\Lambda'_R} = \left\{w; w \text{ is a } \Lambda'_R\text{-MCS}\right\}$, where $R_\Box, \texttt{Choice}, \{\approx_\alpha\}_{\alpha \in Ags}, \left\{R^I_\alpha\right\}_{\alpha \in Ags}$, and $\mathcal{V}$ are defined just as in Chapter 5's Definition D.22 (p. 257), and where $\texttt{Value}_{\mathcal{O}}$ and $\texttt{Value}_{\mathcal{S}}$ are defined as follows: for $\alpha \in Ags$ and $w \in W^{\Lambda}$, one first defines $\Sigma^w_\alpha := \{[\alpha]\varphi; \odot[\alpha]\varphi \in w\}$ and $\Gamma^w_\alpha := \{K_\alpha\varphi; \odot_{\mathcal{S}}[\alpha]\varphi \in w\}$. Then, taking $\Sigma^w = \bigcup_{\alpha\in Ags} \Sigma^w_\alpha$ and $\Gamma^w = \bigcup_{\alpha\in Ags}$, the deontic functions are given by

$$\begin{aligned}\texttt{Value}_{\mathcal{O}}(w) &= \begin{cases} 1 \text{ iff } \Sigma^w \subseteq w, \\ 0 \text{ otherwise.}\end{cases} \\ \texttt{Value}_{\mathcal{S}}(w) &= \begin{cases} 1 \text{ iff } \Gamma^w \subseteq w, \\ 0 \text{ otherwise.}\end{cases}\end{aligned}$$

The canonical structure $\mathcal{M}$ is shown to be a bi-valued Kripke-*kios*-model just as in Chapter 5's Proposition D.23 (p. 257). Then, the so-called *truth lemma* is shown by merging Lemma C.52 (p. 215), Lemma D.25 (p. 259), and Lemma 4 in `https://doi.org/10.48550/arXiv.1903.10577`. This renders completeness with respect to bi-valued Kripke-*kios*-models. □

Proposition E.9 (Completeness w.r.t. bi-valued *kiobt*-models)**.** *The proof system* $\Lambda_R{}'$ *is complete with respect to the class of bi-valued* kiobt-*models.*

Proof. Let φ be a Λ'_R-consistent formula of $\mathcal{L}_{\mathsf{R}}$. Proposition E.8 implies that there exists a bi-valued Kripke-*kios*-model $\mathcal{M}$ and a world w in its domain such that $\mathcal{M}, w \models \varphi$. Proposition E.7 then ensures that the bi-valued *kiobt*-model $\mathcal{M}^T$ associated with $\mathcal{M}$ is such that $\mathcal{M}^T, \langle \overline{w}, h_w\rangle \models \varphi$. □

Therefore, Proposition E.1 and Proposition E.9 imply that the following result, appearing in the main body of the chapter, is true:

Theorem 6.8 (Soundness & Completeness of Λ'_R)**.** *Let* Λ'_R *be the proof system obtained from* Λ_R *by eliminating* (*ConSO*) *in Definition 6.5. Then* Λ'_R *is sound and complete with respect to the class of bi-valued* kiobt-*models.*

Conclusion

> *'You can't bypass nature with logic alone! Logic will presuppose three cases, when there are a million of them! Cut away the whole million, and reduce everything to the one question of comfort... the whole of life's mystery can fit on two printed pages!'*
>
> Fyodor Dostoevsky, *Crime and Punishment*

In the summer of 2019 I heard an eye-opening talk during an artificial intelligence conference. The speaker argued that a path toward constructing ethical AI involves working under an analogy between AI and the human brain. More precisely, the speaker was talking about how we could build hybrid AI systems by dividing their tasks into right-hemisphere tasks and left-hemisphere ones. To clarify, recall that there is a popular and widespread belief that, while the right half of the human brain is used in emotional, creative tasks—where inventiveness and adaptability is required—the left half is used in analytical, logical tasks—where rules and order are required.[26] Thus, the point of the talk was that, to build ethical AI, a promising strategy is to mix-and-match different techniques: while sub-symbolic techniques based on learning algorithms can be used to optimize unsupervised decision-making (right-hemisphere tasks), symbolic techniques based on Logic can be used to harness these decisions and constrain them to specific codes of conduct (left-hemisphere tasks).

I want to conclude this thesis by revisiting a discussion, that began in Chapter 1, regarding the applicability of logic-based frameworks in the development of responsible AI. In particular, I want to position the contents of the thesis as realistic first steps toward the far-reaching goal of developing hybrid ethical AI systems.

[26] This belief started in the 1960's, with the study of patients who had undergone split-brain surgery in which the main commissures connecting the two hemispheres were cut as a means of controlling epilepsy. Testing of each disconnected hemisphere revealed the left to be specialized for language and the right for emotional and nonverbal functions (Corballis, 2014; Sperry, 1982).

Recall from Chapter 1's Section 1.3 that the underlying motivation for my work is to aid in the construction of formal frameworks for performing computational checks on responsibilities of AI systems, just as intended by the research project *REINS* (REsponsible Intelligent Systems) (Broersen, 2014b). It is precisely because of projects like this that I envision the real-life possibility of building ethical AI under a similar strategy to the one described by the conference speaker.

Let me elaborate on such a possibility. Over the last decade, most AI systems have been designed and manufactured using sub-symbolic techniques based on machine learning. Thus, some people have become skeptical about the role of symbolic techniques—or *good old fashioned AI* (GOFAI) (Haugeland, 1985)—in the actual engineering of artificial intelligence. Now, to disambiguate the dichotomy between symbolic and sub-symbolic AI, let me first review what is usually meant when one uses these terms:

- Sub-symbolic AI refers to a variety of methods that involve the following processes: handling large amounts of raw data, performing calculations on this data, recognizing patterns in it (thus *learning* from it), and making predictions/decisions with an implicit, bottom-up kind of intelligence. The traditional paradigm in sub-symbolic AI is machine learning.
- In contrast, symbolic AI refers to techniques that imply explicit models of knowledge and action. The idea is that intelligence can be rendered through the rule-based manipulation of symbols that encode those notions of knowledge and action, where these symbols are given within particular logics.

Although skepticism about GOFAI in AI manufacturing translates into skepticism about the applicability of formal theories of responsibility in the development of ethical AI, there are two main arguments—that are related to one another—to sustain the claim that such theories are more relevant than ever. One has to do with the resurgence of symbolic approaches in foundations & verification of AI, and the other has to do with the suitability of these symbolic—logic-based—approaches for creating hybrid explainable & ethical AI:

1. *AI foundations & verification*: although sub-symbolic techniques have dominated the latest advancements in artificial intelligence, Calegari et al. (2020) recently stated that symbolic AI is re-gaining momentum, specially in the context of tackling one of the biggest problems with sub-symbolic methods: the inability to explain why a system made a decision. For sub-symbolic AI, making a system's underlying decision process completely understandable to human beings is incredibly difficult—if not impossible. Indeed, it

is well-known that sub-symbolic techniques have complex mathematical foundations, on the one hand, and that they admit a very low model interpretability, on the other (where 'model interpretability' refers to a measure of how easy it is for a human being to comprehend the predictions of a system's models). Thus, researchers are turning their attention to tactics that might help in dealing with these problems, something that has led to an important revival of logic-based methods. The reason is that in logic-based methods there is a clear formulation both of systems' models and of the rules that govern systems' decision making. The fields of *knowledge representation*, *logic-based reasoning*, and *formal verification*, among others, all involve perspectives that have their origin in Logic, and they are all starting to be seriously exploited in the design of AI.[27] For instance, according to Calegari et al. (2020), some areas of recent application are formalization & verification, cognitive agents, healthcare & well-being, law & governance, education planning, task allocation, and robotics & control.

2. *Hybrid explainable & ethical AI*: since sub-symbolic AI is considerably opaque, many people—particularly non-experts—do not trust it. This prevents the use of AI systems in activities where autonomous and intelligent decision-making admits moral (and legal) consequences. The idea is that "it is often not sufficient for intelligent systems to produce bare decisions—they must also be *explained*, as ethical and legal issues may arise" (Calegari et al., 2020, p. 15, emphasis in original). Moreover, the use of sub-symbolic methods

[27] The fields of AI that are highly influenced by Logic can be described as follows, following Markman (2013), Calegari et al. (2020), and (Bjesse, 2005):

- *Knowledge representation* is the field of AI that aims to model information about the world in such a way that a computer system can functionally use it. Since virtually nobody argues with the assumption that reasoning requires knowledge, the questions addressed by this field have been fundamental since the early days of AI. Many kinds of (logic-based) knowledge representation systems have been proposed over the years. They mostly rely on description logics and modal logics, used to respectively represent terminological knowledge and time-dependent or subjective knowledge.
- *Logic-based reasoning* includes a number of methods for modelling reasoning. The intuition that underlies all these methods is that formalizing commonsense reasoning comes in handy when building intelligent systems. The following paradigms can be seen to fall under logic-based reasoning: *deduction* (which is the basis of automated theorem-proving and logic programming, for instance), *induction* (which is the basis of inductive model checkers, for instance), *abduction* (which is used in the verification of compliance of specific properties), *non-monotonic reasoning* (which underlies default and defeasible reasoning), and *cognitive-agent architectures* (for which the typical logics are *BDI* or *beliefs-desires-intentions*—see Chapter 5's Section 5.2).
- In the context of hardware and software systems, *formal verification* refers to the formulation within a particular logic of a computer algorithm that underlies any such system. The goal is to check whether the algorithm correctly satisfies a given specification, encoded as a formula of the logic's language. In such a practice, two traditions stand out: (a) *automated theorem-proving*,

implies another dangerous threat: *bias of training data*. If the data upon which a machine-learning algorithm is trained was biased, then its predictions/decisions will unavoidably lead to prejudice (both in the colloquial and the legal connotations of the word 'prejudice'). This is highly problematic, to say the least.[28] Thus, taking measures against these problems is of the utmost importance if we are to allow that AI systems perform tasks in healthcare, governance, and the financial and military industries, for instance. To clarify, reliability and accountability are crucial to prevent catastrophes in all these spheres, and both the lack of transparency in learning algorithms and the possible biases of training data make it difficult to trust AI and to ascertain who should be held liable for an undesirable outcome of an AI system's decision making.[29]

One of the measures taken against these two problems, then, is the integration of sub-symbolic and symbolic AI in what is known as *hybrid intelligent systems* (Corchado et al., 2012; Medsker, 2012). To be more precise, hybrid intelligent systems are models of AI that combine sub-symbolic and symbolic approaches, used both at the level of design/construction and at the level of verification. At the level of design, researchers seek after an explicit integration of symbolic and sub-symbolic models—just as in *neuro-fuzzy systems* (Nauck, Klawonn, & Kruse, 1997; Wu, Zhang, & Lu, 2011) and *neural-symbolic computing* (A. Garcez et al., 2019; A. d. Garcez et al., 2022).[30] At the level of verification, researchers have paid attention to a process known as post-hoc *extraction*, where symbolic knowledge is drawn out from trained numeric predictors in the form of rules that could explain

that refers to the use of a computer program to yield the theorems of a given proof system; and (b) *model checking*, that refers to a series of methods used to assess whether a mathematical model of a computer system satisfies a formula that encodes some specification.

[28] A typical, real-life example of this danger is given by the employment of the software COMPAS (Correctional Offender Management Profiling for Alternative Sanctions) to predict a person's likelihood of becoming a repeat offender. In 2016 it was shown that the algorithm, allegedly based on machine learning, displayed a double racial bias, one in favor of white defendants and one against black defendants (see Flores, Bechtel, & Lowenkamp, 2016; Mehrabi, Morstatter, Saxena, Lerman, & Galstyan, 2021). Although COMPAS is most probably a machine learning algorithm, the model underlying this tool is closed-source and unknown to the public, according to `https://afraenkel.github.io/fairness-book/intro.html`.

[29] Accountability is usually thought of as the state of being liable for an action, meaning being answerable to society for having brought about a state of affairs.

[30] As its name implies, *neuro-fuzzy systems* aim at a synergy of fuzzy and neural systems, combining human-like imprecise reasoning with neural-network learning. In turn, *neural-symbolic computing* integrates robust learning with logic-based reasoning and with a symbolic simplification of artificial neural networks, combining the benefits of meta-heuristics, neural networks, and logic programming.

> their behavior (Andrews, Diederich, & Tickle, 1995; Guidotti et al., 2018).[31] To be sure, hybrid intelligent systems are particularly relevant in *explainable artificial intelligence* (XAI) (Gunning, 2017), a field that aims precisely at exploiting symbolic descriptions in the explanation of the internal functioning of sub-symbolic AI, thus making it more interpretable in the eyes of human beings and better suited to tasks with ethical implications (Arrieta et al., 2020).

So how, then, could this thesis's logics be used to design, verify, and explain ethical AI?

First of all, observe that the frameworks in Chapters 2–6 are already part of the symbolic-AI tradition. More precisely, and as mentioned in Chapter 1 (p. 17), they fall into the category of *agent-based* symbolic AI (see, for instance, Russell & Norvig, 1995; Shoham, 1993; Wooldridge & Jennings, 1995). The *agents* appearing in every stit theory of this thesis are assumed to be entities that reason (they are rational/cognitive), that make their own choices independently (they are autonomous), that interact with other such entities (they are interactive), that perceive the environment and react to it (they are reactive), and that take action in order to achieve their goals and intentions (they are proactive) (see, for instance, Cardoso & Ferrando, 2021; Molina, 2020, for a discussion of these properties of intelligent agents). Indeed, Chapter 2 explicitly formalizes agents in branching time and their taking action, Chapter 3 adds traditional epistemic notions (knowledge and belief) to formalize aspects of such agents' reasoning, Chapter 4

[31] According to Craven (1996), post-hoc *rule extraction* refers to a variety of algorithms that, given a trained neural network and the data on which it was trained, produce a description of the network's hypothesis that is comprehensible and that closely approximates the network's predictive behavior. In other words, rule extraction helps to explain the process of how the network comes to a final decision (Hailesilassie, 2016). The literature has somewhat agreed on a taxonomy for such algorithms (see, for instance Andrews et al., 1995; Guidotti et al., 2018; Hailesilassie, 2016). Such a taxonomy categorizes the algorithms according to the following dimensions: the *expressivity* of the rendered rules, the *translucency* of the algorithm, the *adequacy and portability* of the algorithm, the *quality* of the rendered rules, and the *complexity* of the algorithm (see Andrews et al., 1995, for details). Two dimensions stand out: the expressivity dimension and the translucency dimension. In the expressivity dimension, the most prominent categories in the literature include (a) *if-then rules* (where the rendered rules are of the form 'if input X meets condition A, then output Y will be labelled by class B'), (b) *m-of-n rules* (which involves rules of the form 'if m of the following n antecedents are true, then output Y will be labelled by class B,' (c) *decision trees* (composed by if-then rules running through branches stemming from a root (see Guidotti et al., 2018, p. 376, for a simple example)), and (d) *fuzzy rules* (which use membership functions to deal with partial truths of the form 'if x is low and y is high, then z is medium,' where low, high, and medium are fuzzy sets with corresponding membership functions). In the translucency dimension, the most prominent categories include (a) the *decompositional* approach (where the focus is on extracting rules at the level of individual—both hidden and output—neurons), (b) the *pedagogical* approach (where a neural network is treated as a black box and the goal is to find the whole network's output for a corresponding input), and (c) the *eclectic* approach (which combines both decompositional and pedagogical techniques).

adds obligations to formalize a particular, deontic aspect of the agents' relation with the environment, Chapter 5 adds intentionality to formalize their proactivity, and Chapter 6 merges the previous chapters into a framework to analyze their responsibility. Therefore, the models and languages of these chapters are naturally connected to those of logics for multi-agent systems (see Chapter 2's Subsection 2.4.2) and of *BDI* logics (see Chapter 5's Section 5.2). This means that my study is potentially applicable, both semantically and syntactically, in laying down actual AI foundations, namely through its feasible use in areas such as *knowledge representation* and *logic-based reasoning*.

Similarly, the frameworks in Chapters 2–6 have current relevance in the contexts of *formal verification*. On the one hand, the models can potentially be used both in *model checking* and in post-hoc *extraction* (from a sub-symbolic AI system); on the other, the proof systems can be used in *automated theorem-proving*. Below, I discuss further these paths of application.

As for model checking, observe that the potential use of my models implies two important processes: (a) one would need to either design or interpret an AI system so that its behavior can be seen as based on the models (where the case of interpreting refers to explaining an already existing sub-symbolic system); and (b) one would need to either *implement* the chapters' logics into model checkers (either new or existing ones). Thus, for a formal specification, one could use computer programs to test whether the explanatory formalization of the system (in terms of a logic's model) meets it or not. As for extraction, the potential use of my models involves (a) interpreting an AI system so as to be based on said models, (b) implementing a rule-extraction algorithm that would extract rules from the AI system, and (c) developing translations that would map the extracted rules to formulas that can be evaluated on the models.

In the specific context of building and/or verifying ethical AI, Chapter 6's logic proves useful. For instance, recall the example of the COMPAS algorithm, a software that is employed to predict the likelihood of repeat-offenses (Footnote 28). Assume that (a) we have managed to *interpret* COMPAS as a stit-theoretic agent that—relative to the performance of a given task—has certain actions, knowledge, beliefs, intentions, and obligations, that (b) we have managed to *implement* a rule-extraction algorithm that explains COMPAS's decision-making (see Footnote 31), and that (c) we have translated the extracted rules into Chapter 6's intentional epistemic act-utilitarian stit theory (*IEAUST*). Thus, we can check whether COMPAS's learning-motored choices deviate from a previously established obligation of avoiding racial bias.[32] Furthermore, one could also check what mode of respon-

[32] Of course, it may be hard to logically define 'avoiding racial bias,' generally speaking, and within stit theory, in particular.

sibility (for making particular predictions), if any, can be assigned to COMPAS. Indeed, this strategy evokes the conference speaker's ideas that I mentioned at the beginning of this conclusion: suppose that we design a system by coupling the COMPAS algorithm with the imagined rule-extraction algorithm and with a set of particular ought-to-do's (such as avoiding racial bias). Whenever an ought-to-do is violated by a learning-motored choice, then our system must report such a violation and either train itself again or ask for new training data. Thus, the learning task can be seen as a right-hemisphere process, and the verification of compliance with respect to the ought-to-do's can be seen as a left-hemisphere process.

For another toy example at the level of verification, suppose that we want to check whether a self-driving car knows that it ought to not run over a dog. To check for this, suppose that we have built a model—similar to Chapter 4's *eaubt*-models, for instance—that *adequately* represents the self-driving car as a stit-theoretic agent facing a specific decision context with options for its navigation.[33] Suppose further that we have managed to implement Chapter 4's epistemic act-utilitarian stit theory (*EAUST*) into a model checker. Then we can verify whether and at which states the self-driving car possesses the knowledge that it ought to not run over a dog by testing whether the model satisfies (validates, perhaps) formulas of the forms $K_{car} \odot_{car} \neg r$ and $K_{car} \odot^{S}_{car} \neg r$ (where r stands for the proposition 'the dog has been run over'). The reason is that $K_{car} \odot_{car} \neg r$, resp. $K_{car} \odot^{S}_{car} \neg r$, holds at a state iff at that state agent *car* knows that it objectively, resp. subjectively, ought to not run over the dog.

Indeed, the REINS project itself (the thesis's motivation, p. 16) intended to do implementations of this kind. Originally, the last stage of the project involved developing translations from logic-based formalisms representing graded responsibilities—exactly like Chapter 6's—to formalisms for which model checkers already exist, with the goal of performing computational checks on responsibilities. As argued by Broersen (2014b), a first possibility for investigation would have been model checker Mocha (Alur et al., 1998), that checks formulas of alternating-time temporal logic (*ATL*) (see Chapter 2's Subsection 2.4.2). Unfortunately, the project did not arrive to this stage, but this only implies that there are still worthy opportunities for future work.

As for automated theorem-proving, observe that the proof systems in Chapters 2–6 can be used as background axiomatic systems against which to test for specific deductions and/or provable formulas using automated theorem-provers.

[33] Of course, a measure of the adequacy of representations is very important. One could use criteria for evaluation (of the adequacy of an AI system's representation within a logic-based model) similar to the ones proposed by Andrews et al. (1995) to check for the quality of rule-extraction algorithms, i.e., test the *accuracy*, *fidelity*, *consistency*, and *comprehensibility* of the representation at hand.

These logics are particularly relevant in the context of verifying ethical AI. For instance, again consider the self-driving car example. Suppose that, for a given navigation-scenario, we have managed to embody agent *car* as a set Σ of formulas of $\mathcal{L}_{\mathsf{KO}}$—the language of Chapter 4's *EAUST*—in terms of *car*'s available choices and abilities. Suppose further that we have managed to characterize *car*'s knowledge-base (for the given navigation-scenario) with a set Γ of formulas of $\mathcal{L}_{\mathsf{KO}}$. Then we can test whether the car knows that it subjectively ought to not run over a dog by using a theorem-prover to see whether $\Sigma \cup \Gamma \vdash_{\Lambda_S} K_{car} \odot^{S}_{car} \neg r$.

Of course, complexity-related issues play a role in the success of hypothetical theorem-provers like the one in the example above. First of all, the logics in Chapters 3–6 might not be decidable, and thus a theorem-prover would never terminate when testing the provability of certain formulas. As a response to this first concern, it is important to mention that existing results suggest that the logics in Chapters 3–6 are likely to be decidable. For *EXST*—the logic in Chapter 3—one can probably adapt Payette's (2014) method to prove its decidability. For *EAUST*, *IEST*, and *IEAUST*—the logics in Chapters 4, 5, and 6, respectively—the results in my joint paper with Jan Broersen (Abarca & Broersen, 2021a), themselves extending Murakami's (2004) proof of decidability for *AUST*, can in principle be adapted to render decidability via finite-model property.[34]

Now, even if the logics were shown to be decidable, a second obstacle is that the provability problem might still be too hard, complexity-wise, and that current theorem-provers could turn out to be inefficient. An already existing paradigm, then, brings hope in this respect. Arkoudas et al. (2005) presented a natural-deduction calculus for the axiomatization that Murakami (2004) gave to Horty's *AUST* (the sound, complete, and decidable proof system that I discussed in Chapter 4's Proposition C.38). The authors encoded this natural-deduction calculus in an interactive theorem-prover named Athena, with which they successfully tested the provability of *AUST* formulas. Thus, Arkoudas et al. proved that Horty's (2001) seminal theory of ought-to-do is "AI-friendly" (Bringsjord, Arkoudas, & Bello, 2006, p. 8) and that it is possible to use "mechanized deontic logics" in AI verification (Arkoudas et al., 2005, p. 23). Since the logics in Chapters 4–6 largely rely on *AUST*'s ideas, it is not far-fetched to think that Arkoudas et al.'s implementation could be extended to them.

[34] It is not hard to show that the single-agent versions of the logics in Chapters 4–6 are decidable. One can prove the finite-model property by first showing completeness with respect to Kripke rooted models (rooted with respect to $\approx_\alpha \circ R_\Box$) and then filtrate those models, in an adaptation of Bezhanishvili's (2006, Chapter 6) methods.

All the scenarios described above (AI design, model checking, theorem proving, extraction, etc.) imply a great deal of work, as well as the confluence of interdisciplinary efforts. To sum up possible paths for future work in these respects, consider Figure 6.4 below:

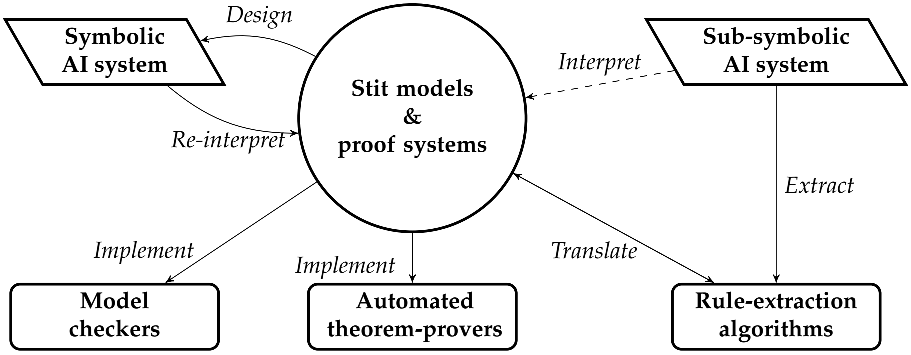


Figure 6.4

Figure 6.4 illustrates how the models and proof systems for the logics in this thesis could be used to (a) design symbolic AI systems and interpret existing sub-symbolic AI systems, (b) implement said logics into model checkers and/or automated theorem-provers, and (c) find translations of rules extracted from sub-symbolic AI systems. Under such a scheme, the nature of these logics should prove them valuable in the design and verification of responsible AI. In the particular case of mix-and-matching sub-symbolic and symbolic approaches, such a scheme is reminiscent of the strategy advocated by the conference speaker: let the learning be done according to usual machine-learning methods, and verify (constrain, or harness) such a learning with logic-based formalisms.

As implied by the present discussions, AI developers face big challenges in the construction of systems that are expected to make decisions with moral consequences. Looking for ways to tackle this challenge, the field of machine ethics has seen a quick growth in recent years. Thus, questions concerning the responsibility of autonomous intelligent agents have become very important. These questions can be categorized in two main trends: (1) conceptual questions about the ontology and essential components of the notion of responsibility, and (2) technical questions that revolve around the implementation of such a notion in AI. This thesis attempted to provide possible answers in both categories. Conceptually, the logics imply an extension of Horty's (2001) *AUST* with epistemic and intentional attitudes (according to the operational definition and decomposition of responsibility proposed in Chapter 1, p. 3). Technically, there are two main contributions that lay groundwork for viable efforts of implementation: on the one hand, the logics offer expressive models of agency, knowledge, belief, intentions, ought-to-do, and responsibilities; on the other hand, the logics admit sound and complete proof systems.

On p. 3 I wrote that this thesis aimed to build a formal theory of responsibility, that the main tool used toward this aim would be Logic, and that the underlying motivation was to provide theoretical foundations for symbolic techniques in the development of ethical AI. Whether and to what extent the goal has been met is an appreciation that I ultimately leave to the reader. However, what is clear to me—after writing all these pages—is that the attempt is already a contribution inasmuch as it opens roads for interesting further research, as argued by this conclusion.

In Dostoevsky's *Crime and Punishment*, the character Razumikhin says that one "can't bypass nature with Logic alone," that Logic "will presuppose three cases, when there are a million of them." I truly believe that there is no arguing that. However, I also believe that, rather than being meant to bypass nature, Logic itself belongs to (human) nature and can help us comprehend it and better our interactions with and within it. This belief is the driving force behind the present work, and I surely hope that something somewhere in all these pages can evoke such a belief in some reader, too.

References

Abarca, A. I. R., & Broersen, J. (2019). A logic of objective and subjective oughts. In *European conference on logics in artificial intelligence* (pp. 629–641).

Abarca, A. I. R., & Broersen, J. (2021a). A deontic stit logic based on beliefs and expected utility. *Electronic Proceedings in Theoretical Computer Science*, *335*, 281–294. Retrieved from `https://doi.org/10.4204%2Feptcs.335.27`

Abarca, A. I. R., & Broersen, J. (2021b). Stit semantics for epistemic notions based on information disclosure in interactive settings. *Journal of Logical and Algebraic Methods in Programming*, *123*, 100708. Retrieved from `https://www.sciencedirect.com/science/article/pii/S2352220821000717`

Abarca, A. I. R., & Broersen, J. (2022). A stit logic of responsibility. In *Proceedings of the 21st international conference on autonomous agents and multiagent systems* (pp. 1717–1719).

Abarca, A. I. R., & Broersen, J. (2023). A stit logic of intentionality. In C. Areces & D. Costa (Eds.), *Dynamic Logic. New Trends and Applications* (pp. 125–153). Springer.

Ågotnes, T. (2006). Action and knowledge in alternating-time temporal logic. *Synthese*, *149*(2), 375–407.

Ågotnes, T., Goranko, V., Jamroga, W., & Wooldridge, M. (2015). Knowledge and ability. In H. van Ditmarsch, J. Halpern, W. van der Hoek, & B. Kooi (Eds.), *Handbook of Epistemic Logic* (pp. 543–589). College Publications.

Ågotnes, T., van der Hoek, W., Rodríguez-Aguilar, J. A., Sierra, C., & Wooldridge, M. J. (2007). On the logic of normative systems. In *IJCAI* (Vol. 7, pp. 1175–1180).

Ahmed, A. (2018). *Newcomb's problem*. Cambridge University Press.

Alur, R., Henzinger, T. A., & Kupferman, O. (1997). Alternating-time temporal logic. In *International Symposium on Compositionality* (pp. 23–60).

Alur, R., Henzinger, T. A., & Kupferman, O. (2002). Alternating-time temporal logic. *Journal of the ACM*, *49*(5), 672–713.

Alur, R., Henzinger, T. A., Mang, F. Y., Qadeer, S., Rajamani, S. K., & Tasiran, S. (1998). Mocha: Modularity in model checking. In *International Conference on Computer Aided Verification* (pp. 521–525).

Anderson, A. R. (1956). *The formal analysis of normative systems*. New Haven, CT, USA: Yale University, International Laboratory, Sociology Dept.

Andrews, R., Diederich, J., & Tickle, A. B. (1995). Survey and critique of techniques for extracting rules from trained artificial neural networks. *Knowledge-Based Systems*, *8*(6), 373–389.

Anscombe, G. E. M. (1963). *Intention*. Cambridge: Harvard University Press.

Arkoudas, K., Bringsjord, S., & Bello, P. (2005). Toward ethical robots via mechanized deontic logic. In *AAAI Fall Symposium on Machine Ethics* (pp. 17–23).

Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., ... Benjamins, R. (2020). Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, *58*, 82–115.

Aumann, R. J. (1999). Interactive epistemology II: probability. *International Journal of Game Theory*, *28*(3), 301–314.

Aumann, R. J., & Dreze, J. H. (2008). Rational expectations in games. *American Economic Review*, *98*(1), 72–86.

Aune, B. (1977). *Reason and action* (Vol. 9). Springer Science & Business Media.

Austin, J. L. (1961). *Philosophical papers*. Oxford University Press.

Balbiani, P., Herzig, A., & Troquard, N. (2008). Alternative axiomatics and complexity of deliberative stit theories. *Journal of Philosophical Logic*, *37*(4), 387–406.

Ballarin, R. (2017). Modern Origins of Modal Logic. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Summer 2017 ed.). Metaphysics Research Lab, Stanford University. `https://plato.stanford.edu/archives/sum2017/entries/logic-modal-origins/`.

Baltag, A., Bezhanishvili, N., Özgün, A., & Smets, S. (2015). The topology of full and weak belief. In *International Tbilisi Symposium on Logic, Language, and Computation* (pp. 205–228).

Baltag, A., Bezhanishvili, N., Özgün, A., & Smets, S. (2016). Justified belief and the topology of evidence. In *International Workshop on Logic, Language, Information, and Computation* (pp. 83–103).

Baltag, A., & Renne, B. (2016). Dynamic Epistemic Logic. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Winter 2016 ed.). Metaphysics Research Lab, Stanford University. `https://plato.stanford.edu/archives/win2016/entries/dynamic-epistemic/`.

Baltag, A., & Smets, S. (2006). Conditional doxastic models: A qualitative approach to dynamic belief revision. *Electronic Notes in Theoretical Computer Science*, *165*, 5–21.

Baltag, A., & Smets, S. (2008). Probabilistic dynamic belief revision. *Synthese*, *165*(2), 179.

Bartha, P. (2014). Decisions in branching time. In *Nuel Belnap on Indeterminism and Free Action* (pp. 29–56). Springer.

Barwise, J. (1989). On the model theory of common knowledge. *The Situation in Logic*, *17*, 201–220.

Baskent, C., Loohuis, L. O., & Parikh, R. (2012). On knowledge and obligation. *Episteme*, *9*(2), 171-188.

Belnap, N., & Perloff, M. (1988). Seeing to it that: a canonical form for agentives. *Theoria*, *54*(3), 175–199.

Belnap, N., Perloff, M., & Xu, M. (2001). *Facing the future: agents and choices in our indeterminist world*. Oxford University Press.

Bentzen, M. M. (2012). *Stit, iit, and deontic logic for action types* (Unpublished doctoral dissertation). University of Amsterdam.

Bergström, L. (1966). *The alternatives and consequences of actions*. Göteborg, Almqvist & Wiksell.

Bezhanishvili, N. (2006). *Lattices of intermediate and cylindric modal logics* (Unpublished doctoral dissertation). University of Amsterdam.

Bjesse, P. (2005). What is formal verification? *ACM SIGDA Newsletter*, *35*(24), 1–es.

Bjorndahl, A., Halpern, J. Y., & Pass, R. (2011). Reasoning about justified belief. In *Proceedings of the 13th Conference on Theoretical Aspects of Rationality and Knowledge* (pp. 221–227).

Bjorndahl, A., Halpern, J. Y., & Pass, R. (2017). Reasoning about rationality. *Games and Economic Behavior*, *104*, 146–164.

Blackburn, P., De Rijke, M., & Venema, Y. (2002). *Modal logic* (Vol. 53). Cambridge University Press.

Board, O. (2004). Dynamic interactive epistemology. *Games and Economic Behavior*, *49*(1), 49–80.

Boella, G., van der Torre, L., & Verhagen, H. (2006). Introduction to normative multiagent systems. *Computational & Mathematical Organization Theory*, *12*(2-3), 71–79.

Bonanno, G. (2004). Memory and perfect recall in extensive games. *Games and Economic Behavior*, *47*(2), 237–256.

Bratman, M. (1984). Two faces of intention. *The Philosophical Review*, *93*(3), 375–405.

Bratman, M. (1987). *Intention, plans, and practical reason*. Cambridge: Cambridge, MA: Harvard University Press.

Bratman, M. (2013). *Shared agency: A planning theory of acting together*. Oxford University Press.

Bringsjord, S., Arkoudas, K., & Bello, P. (2006). Toward a general logicist methodology for engineering ethically correct robots. *IEEE Intelligent Systems*, *21*(4), 38–44.

Broersen, J. (2008a). A complete stit logic for knowledge and action, and some of its applications. In *International Workshop on Declarative Agent Languages and Technologies* (pp. 47–59).

Broersen, J. (2008b). A logical analysis of the interaction between 'obligation-to-do' and 'knowingly doing'. In *International Conference on Deontic Logic in Computer Science* (pp. 140–154).

Broersen, J. (2011a). Deontic epistemic stit logic distinguishing modes of mens rea. *Journal of Applied Logic*, *9*(2), 137–152.

Broersen, J. (2011b). Making a start with the stit logic analysis of intentional action. *Journal of Philosophical Logic*, *40*(4), 499–530.

Broersen, J. (2011c). Probabilistic stit logic. In *European Conference on Symbolic and Quantitative Approaches to Reasoning and Uncertainty* (pp. 521–531).

Broersen, J. (2014a). On the reconciliation of logics of agency and logics of event types. In *Krister Segerberg on Logic of Actions* (pp. 41–59). Springer.

Broersen, J. (2014b). Responsible intelligent systems. *KI-Künstliche Intelligenz*, *28*(3), 209–214.

Broersen, J., & Abarca, A. I. R. (2018a). Formalising oughts and practical knowledge without resorting to action types. In *Proceedings of the 17th International Conference on Autonomous Agents and Multiagent Systems* (pp. 1877–1879).

Broersen, J., & Abarca, A. I. R. (2018b). Knowledge and subjective oughts in stit logic. In *Proceedings of DEON* (pp. 51–69).

Broersen, J., Herzig, A., & Troquard, N. (2006a). Embedding alternating-time temporal logic in strategic logic of agency. *Journal of Logic and Computation*, *16*(5), 559–578.

Broersen, J., Herzig, A., & Troquard, N. (2006b). From coalition logic to stit. *Electronic Notes in Theoretical Computer Science*, *157*(4), 23–35.

Brown, M. A. (1988). On the logic of ability. *Journal of Philosophical Logic*, *17*(1), 1–26.

Burgess, J. P. (1979). Logic and time. *The Journal of Symbolic logic*, *44*(4), 566–582.

Calegari, R., Ciatto, G., Denti, E., & Omicini, A. (2020). Logic-based technologies for intelligent systems: State of the art and perspectives. *Information*, *11*(3), 167.

Canavotto, I. (2020). *Where responsibility takes you: Logics of agency, counterfactuals and norms* (Unpublished doctoral dissertation). Institute for Logic, Language and Computation.

Cardoso, R. C., & Ferrando, A. (2021). A review of agent-based programming for multi-agent systems. *Computers*, *10*(2), 16.

Castañeda, H.-N. (1968). A problem for utilitarianism. *Analysis*, *28*(4), 141–142.

Chellas, B. F. (1969). *The logical form of imperatives*. Stanford University.

Chellas, B. F. (1980). *Modal logic: an introduction*. Cambridge University press.

Chellas, B. F. (1992). Time and modality in the logic of agency. *Studia Logica*, *51*(3-4), 485–517.

Chisholm, R. M. (1963). Contrary-to-duty imperatives and deontic logic. *Analysis*, *24*(2), 33–36.

Chisholm, R. M. (1964). The ethics of requirement. *American Philosophical Quarterly*, *1*(2), 147–153.

Ciuni, R., & Horty, J. F. (2014). Stit logics, games, knowledge, and freedom. In *Johan van Benthem on Logic and Information Dynamics* (pp. 631–656). Springer.

Ciuni, R., & Lorini, E. (2017). Comparing semantics for temporal stit logic. *Logique et Analyse*, *61*(243), 299–339.

Coeckelbergh, M. (2020). Artificial intelligence, responsibility attribution, and a relational justification of explainability. *Science and Engineering Ethics*, *26*(4), 2051–2068.

Cohen, P. R., & Levesque, H. J. (1990). Intention is choice with commitment. *Artificial intelligence*, *42*(2-3), 213–261.

Corballis, M. C. (2014). Left brain, right brain: facts and fantasies. *PLoS biology*, *12*(1). Retrieved from `https://doi.org/10.1371/journal.pbio.1001767`

Corchado, E., Snasel, V., Abraham, A., Woźniak, M., Grana, M., & Cho, S. (2012). *Hybrid artificial intelligent systems*. Springer.

Craven, M. W. (1996). *Extracting comprehensible models from trained neural networks* (Unpublished doctoral dissertation). The University of Wisconsin-Madison.

Crisp, R. (2014). *Aristotle: Nicomachean ethics*. Cambridge University Press.

Davidson, D. (1963). Actions, reasons, and causes. *The Journal of Philosophy*, *60*(23), 685–700. (reprinted in 1980, pp. 3–20)

Davidson, D. (1978). Intending. In *Philosophy of History and Action* (pp. 41–60). Springer.

Davidson, D. (1980). *Essays on actions and events*. Oxford: Clarendon Press.

Dubber, M. D. (2002). *Model penal code*. New York: Foundation Press.

Duijf, H. (2018). *Let's do it!: Collective responsibility, joint action, and participation* (Unpublished doctoral dissertation). Utrecht University.

Duijf, H. (2022). *Logic of responsibility voids*. Springer.

Duijf, H., Broersen, J., Kuncová, A., & Abarca, A. I. R. (2021). Doing without action types. *The Review of Symbolic Logic*, *14*(2), 380–410.

Dunin-Keplicz, B., & Verbrugge, R. (2002). Collective intentions. *Fundamenta Informaticae*, *51*(3), 271–295.

Eells, E. (1982). *Rational decision and causality*. Cambridge University Press.

Engelking, R. (1989). *General topology*. Heldermann.

Enqvist, S. (2005). *Completeness in modal logic.* `https://lup.lub.lu.se/luur/download?func=downloadFile&recordOId=1328154&fileOId=1328155`. (unpublished)

Fagin, R., Moses, Y., Halpern, J. Y., & Vardi, M. Y. (1995). *Reasoning about knowledge*. MIT press.

Falvey, K. (2000). Knowledge in intention. *Philosophical Studies: An International Journal for Philosophy in the Analytic Tradition*, *99*(1), 21–44.

Fantl, J. (2008). Knowing-how and knowing-that. *Philosophy Compass*, *3*(3), 451–470.

Fischer, J. M. (1982). Responsibility and control. *The Journal of Philosophy*, *79*(1), 24–40.

Fischer, M. J., & Ladner, R. E. (1979). Propositional dynamic logic of regular programs. *Journal of Computer and System Sciences*, *18*(2), 194–211.

Flasiński, M. (2016). Symbolic artificial intelligence. In *Introduction to Artificial Intelligence* (pp. 15–22). Springer.

Flores, A. W., Bechtel, K., & Lowenkamp, C. T. (2016). False positives, false negatives, and false analyses: A rejoinder to machine bias: There's software used across the country to predict future criminals. and it's biased against blacks. *Fed. Probation*, *80*, 38.

Føllesdal, D., & Hilpinen, R. (1970). Deontic logic: An introduction. In *Deontic Logic: Introductory and systematic readings* (pp. 1–35). Springer.

Frankfurt, H. (2018). Alternate possibilities and moral responsibility. In *Moral Responsibility and Alternative Possibilities* (pp. 17–25). Routledge.

Gabbay, D. M., Hodkinson, I., & Reynolds, M. A. (1994). *Temporal logic: Mathematical foundations and computational aspects*. Oxford University Press (on Demand).

Garcez, A., Gori, M., Lamb, L., Serafini, L., Spranger, M., & Tran, S. (2019). Neural-symbolic computing: An effective methodology for principled integration of machine learning and reasoning. *Journal of Applied Logics*, *6*(4), 611–632.

Garcez, A. d., Bader, S., Bowman, H., Lamb, L. C., de Penning, L., Illuminoo, B., . . . Gerson Zaverucha, C. (2022). Neural-symbolic learning and reasoning: A survey and interpretation. *Neuro-Symbolic Artificial Intelligence: The State of the Art*, *342*, 1.

Garson, J. (2021). Modal Logic. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Summer 2021 ed.). Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/sum2021/entries/logic-modal/.

Geach, P. T. (1991). Whatever happened to deontic logic? In *Logic and Ethics* (pp. 33–48). Springer.

Gerbrandy, J. (1998). Distributed knowledge. *Twendial*, *98*, 111–124.

Gibbard, A., & Harper, W. L. (1978). Counterfactuals and two kinds of expected utility. In A. Hooker, J. J. Leach, & E. F. McClennen (Eds.), *Foundations and Applications of Decision Theory* (pp. 125–162). D. Reidel.

Goranko, V. (2001). Coalition games and alternating temporal logics. In *Theoretical Aspects Of Rationality And Knowledge: Proceedings of the 8th conference on Theoretical aspects of rationality and knowledge* (Vol. 8, pp. 259–272).

Goranko, V., & van Drimmelen, G. (2006). Complete axiomatization and decidability of alternating-time temporal logic. *Theoretical Computer Science*, *353*(1-3), 93–117.

Gorr, M., & Horgan, T. (1982). Intentional and unintentional actions. *Philosophical Studies: An International Journal for Philosophy in the Analytic Tradition*, *41*(2), 251–262.

Green, S. P. (2005). Six senses of strict liability: A plea for formalism. *Appraising Strict Liability*, *1*.

Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., & Pedreschi, D. (2018). A survey of methods for explaining black box models. *ACM computing surveys (CSUR)*, *51*(5), 1–42.

Gunning, D. (2017). Explainable artificial intelligence (XAI). *Defense Advanced Research Projects Agency (DARPA)*, *2*(2).

Hailesilassie, T. (2016). Rule extraction algorithm for deep neural networks: A review. *arXiv preprint arXiv:1610.05267*. Retrieved from https://arxiv.org/abs/1610.05267

Halpern, J. Y., & Fagin, R. (1989). Modelling knowledge and action in distributed systems. *Distributed Computing*, *3*(4), 159–177.

Hansson, B., & Gärdenfors, P. (1973). A guide to intensional semantics. *Modality, Morality and Other Problems of Sense and Nonsense. Essays Dedicated to Sören Halldén*, 151–167.

Harel, D. (1984). Dynamic logic. In *Handbook of Philosophical Logic* (pp. 497–604). Springer.

Harel, D., Kozen, D., & Tiuryn, J. (2001). Dynamic logic. In *Handbook of Philosophical Logic* (pp. 99–217). Springer.

Harsanyi, J. C. (1967). Games with incomplete information played by "Bayesian" players, i–iii part i. the basic model. *Management Science*, *14*(3), 159–182.

Haugeland, J. (1985). *Artificial intelligence: The very idea*. MIT press.

Helbing, D. (2012). Agent-based modeling. In *Social Self-Organization* (pp. 25–70). Springer.

Herzig, A., & Longin, D. (2004). C&L intention revisited. *KR*, *4*, 527–535.

Herzig, A., & Lorini, E. (2010). A dynamic logic of agency I: Stit, capabilities and powers. *Journal of Logic, Language and Information*, *19*(1), 89-121. Retrieved from http://dx.doi.org/10.1007/s10849-009-9105-x

Herzig, A., & Schwarzentruber, F. (2008). Properties of logics of individual and group agency. *Advances in Modal Logic*, *7*, 133–149.

Herzig, A., & Troquard, N. (2006). Knowing how to play: uniform choices in logics of agency. In *Proceedings of the 5th International Joint Conference on Autonomous Agents and Multiagent Systems* (pp. 209–216).

Hilpinen, R., & McNamara, P. (2013). Deontic logic: A historical survey and introduction. *Handbook of Deontic Logic and Normative Systems*, *1*, 3–136.

Hintikka, J. (1962). *Knowledge and belief: An introduction to the logic of the two notions*. Ithaca: Cornell University Press.

Hoek, W. v. d., & Wooldridge, M. (2003). Cooperation, knowledge, and time: Alternating-time temporal epistemic logic and its applications. *Studia Logica*, *75*(1), 125–157.

Horty, J. F. (1989). An alternative stit operator. *Manuscript, Philosophy Department, University of Maryland*, *11*, 88–105.

Horty, J. F. (2001). *Agency and deontic logic*. Oxford University Press.

Horty, J. F. (2019). *Epistemic oughts in stit semantics*. Ann Arbor, MI: Michigan Publishing, University of Michigan Library.

Horty, J. F., & Belnap, N. (1995). The deliberative stit: A study of action, omission, ability, and obligation. *Journal of Philosophical Logic*, *24*(6), 583–644.

Horty, J. F., & Pacuit, E. (2017). Action types in stit semantics. *The Review of Symbolic Logic*, *10*(4), 617–637.

Jamroga, W., & Ågotnes, T. (2007). Constructive knowledge: what agents can achieve under imperfect information. *Journal of Applied Non-Classical Logics*, *17*(4), 423–475.

Jeffrey, R. C. (1965). *The logic of decision*. McGraw-Hill Book Company.

Jennings, N. R. (2000). On agent-based software engineering. *Artificial intelligence*, *117*(2), 277–296.

Jörgensen, E. (1984). 'Ought'. present or past tense? *English Studies*, *65*(6), 550-554. Retrieved from https://doi.org/10.1080/00138388408598359

Kanger, S. (1970). New foundations for ethical theory. In *Deontic logic: Introductory and Systematic Readings* (pp. 36–58). Springer.

Kenny, A. (1976a). Human abilities and dynamic modalities. In *Essays on Explanation and Understanding* (pp. 209–232). Springer.

Kenny, A. (1976b). *Will, freedom, and power*. Blackwell.

Kolmogorov, A. N. (1956). *Foundations of the theory of probability* (Second English Edition, 2018 ed.). Courier Dover Publications.

Kolodny, N., & MacFarlane, J. (2010). Ifs and oughts. *Journal of Philosophy*, *107*(3), 115-43.

Konolige, K., & Pollack, M. E. (1993). A representationalist theory of intention. In *IJCAI* (Vol. 93, pp. 390–395).

Kooi, B., & Tamminga, A. (2008). Moral conflicts between groups of agents. *Journal of Philosophical Logic*, *37*(1), 1–21.

Kripke, S. A. (1959). A completeness theorem in modal logic. *The Journal of Symbolic Logic*, *24*(1), 1–14.

Kripke, S. A. (1963). Semantical analysis of modal logic I. normal modal propositional calculi. *Mathematical Logic Quarterly*, *9*(5-6), 67–96.

Larkin JR, P. J. (2014). Strict liability offenses, incarceration, and the cruel and unusual punishments clause. *Harvard Journal of Law and Public Policy*, *37*(3), 1065.

Lenzen, W. (1979). Epistemologische betrachtungen zu [s4, s5]. *Erkenntnis*, *14*(1), 33–56.

Lindström, S., & Segerberg, K. (2007). Modal logic and philosophy. In P. Blackburn, J. van Benthem, & F. Wolter (Eds.), *Handbook of Modal Logic* (pp. 1149–1214). Amsterdam, the Netherlands: Elsevier.

Lorini, E. (2013). Temporal logic and its application to normative reasoning. *Journal of Applied Non-Classical Logics*, *23*(4), 372–399.

Lorini, E., & Herzig, A. (2008). A logic of intention and attempt. *Synthese*, *163*(1), 45–77.

Lorini, E., Longin, D., & Mayor, E. (2014). A logical analysis of responsibility attribution: emotions, individuals and collectives. *Journal of Logic and Computation*, *24*(6), 1313–1339.

Lorini, E., & Sartor, G. (2016). A stit logic for reasoning about social influence. *Studia Logica*, *104*, 773–812.

Luce, R. D., & Raiffa, H. (1957). *Games and decisions: Introduction and critical survey*. John Wiley and Sons, New York.

Machina, M., & Viscusi, W. K. (2013). *Handbook of the economics of risk and uncertainty*. Newnes.

Marcus, E. (2019). Reconciling practical knowledge with self-deception. *Mind, 128*(512), 1205–1225.

Markman, A. B. (2013). *Knowledge representation*. Psychology Press.

Medsker, L. R. (2012). *Hybrid intelligent systems*. Springer Science & Business Media.

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR), 54*(6), 1–35.

Mendelson, E. (1964). *Introduction to mathematical logic*. Chapman and Hall.

Menzel, C. (2017). Possible Worlds. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Winter 2017 ed.). Metaphysics Research Lab, Stanford University. `https://plato.stanford.edu/archives/win2017/entries/possible-worlds/`.

Meyer, J.-J. C. (1988). 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.

Meyer, J.-J. C., van der Hoek, W., & van Linder, B. (1999). A logical approach to the dynamics of commitments. *Artificial Intelligence, 113*(1-2), 1–40.

Molina, M. (2020). What is an intelligent system? *arXiv preprint arXiv:2009.09083*.

Montague, R. (1970). Universal grammar. *Theoria, 36*(3), 373–398.

Moran, R., & Stone, M. J. (2009). Anscombe on expression of intention. In *New Essays On the Explanation of Action* (pp. 132–168). Springer.

Murakami, Y. (2004). Utilitarian deontic logic. *Advances in Modal Logic, 287*.

Nauck, D., Klawonn, F., & Kruse, R. (1997). *Foundations of neuro-fuzzy systems*. John Wiley & Sons, Inc.

Naumov, P., & Tao, J. (2017). Coalition power in epistemic transition systems. In *Proceedings of the 16th Conference on Autonomous Agents and Multiagent Systems* (pp. 723–731).

Naumov, P., & Tao, J. (2018). Together we know how to achieve: An epistemic logic of know-how. *Artificial Intelligence, 262*, 279–300.

Nilsson, N. J., & Nilsson, N. J. (1998). *Artificial intelligence: a new synthesis*. Morgan Kaufmann.

Nozick, R. (1969). Newcomb's problem and two principles of choice. In *Essays in Honor of Carl G. Hempel* (pp. 114–146). Springer.

Olsen, C. (1969). Knowledge of one's own intentional actions. *The Philosophical Quarterly (1950-), 19*(77), 324–336.

Osborne, M. J., & Rubinstein, A. (1994). *A course in game theory*. MIT press.

Özgün, A. (2017). *Evidence in epistemic logic: a topological perspective* (Unpublished doctoral dissertation). Université de Lorraine.

Pacuit, E. (2007). *Neighborhood semantics for modal logic: An introduction.* (Course Notes for ESSLLI)

Pacuit, E., Parikh, R., & Cogan, E. (2006). The logic of knowledge based obligation. *Knowledge, Rationality and Action a subjournal of Synthese*, *149*(2), 311–341.

Pacuit, E., & Roy, O. (2017). Epistemic Foundations of Game Theory. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy* (Summer 2017 ed.). Metaphysics Research Lab, Stanford University. `https://plato.stanford.edu/archives/sum2017/entries/epistemic-game/`.

Parfit, D. (1988). *What we together do.* (Unpublished manuscript)

Pauly, M. (2001). *Logic for social software* (Unpublished doctoral dissertation). Universiteit van Amsterdam.

Pauly, M. (2002). A modal logic for coalitional power in games. *Journal of logic and computation*, *12*(1), 149–166.

Payette, G. (2012). *Completeness of xpstit logic.* `https://www.academia.edu/1857669/Completeness_of_XPstit_Logic_An_extension_of_Xstit`. (unpublished)

Payette, G. (2014). Decidability of an xstit logic. *Studia Logica*, *102*(3), 577–607.

Pelletier, J. (1977). Formal philosophy. *Metaphilosophy*, *8*(4), 320–341. Retrieved from `http://www.jstor.org/stable/24435424`

Perea, A. (2012). *Epistemic game theory: reasoning and choice*. Cambridge University Press.

Pereira, L. M., & Saptawijaya, A. (2016). *Programming machine ethics* (Vol. 26). Springer.

Perloff, M. (1991). Stit and the language of agency. *Synthese*, *86*(3), 379–408.

Pratt, V. R. (1982). Dynamic logic. In *Studies in Logic and the Foundations of Mathematics* (Vol. 104, pp. 251–261). Elsevier.

Prior, A. N. (1967). *Past, present and future* (Vol. 154). Clarendon Press Oxford.

Prior, A. N., & Prior, N. (1955). *Formal logic*. Oxford University Press.

Ramsey, F. (1926). Truth and probability. reprinted in. *Studies in Subjective Probability*, 61–92.

Rao, A. S., & Georgeff, M. P. (1991). Modeling rational agents within a bdi-architecture. *KR*, *91*, 473–484.

Raz, J. (1999). *Practical reason and norms*. OUP Oxford.

Russell, S. J., & Norvig, P. (1995). *Artificial intelligence: A modern approach*. Prentice-Hall.

Ryle, G. (2009). *The concept of mind*. Routledge.

Savage, L. J. (1951). The theory of statistical decision. *Journal of the American Statistical Association*, *46*(253), 55–67.

Savage, L. J. (1954). *The foundations of statistics*. New York: John Wiley and Sons.

Schlosser, M. (2019). Agency. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Winter 2019 ed.). Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/win2019/entries/agency/.

Schwarzentruber, F. (2012). Complexity results of stit fragments. *Studia Logica, 100*(5), 1001–1045.

Scott, D. (1970). Advice on modal logic. In *Philosophical Problems in Logic* (pp. 143–173). Springer.

Searle, J. R. (1983). *Intentionality: An essay in the philosophy of mind*. Cambridge University Press.

Segerberg, K. (1992). Getting started: Beginnings in the logic of action. *Studia Logica*, *51*(3-4), 347–378.

Segerberg, K., Meyer, J.-J., & Kracht, M. (2017). The Logic of Action. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Fall 2017 ed.). Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/fall2017/entries/logic-action/.

Setiya, K. (2011). Knowledge of Intention. In *Essays on Anscombe's intention* (pp. 170–197). Harvard University Press.

Setiya, K. (2018). Intention. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Fall 2018 ed.). Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/fall2018/entries/intention/.

Shoham, Y. (1993). Agent-oriented programming. *Artificial Intelligence*, *60*(1), 51–92.

Singh, M. P. (1999). Know-how. In A. S. Rao & M. Wooldridge (Eds.), *Foundations of Rational Agency* (pp. 105–132). Springer.

Sperry, R. (1982). Some effects of disconnecting the cerebral hemispheres. *Science, 217*(4566), 1223–1226.

Stalnaker, R. (1991). The problem of logical omniscience, I. *Synthese*, *89*(3), 425–440.

Stalnaker, R. (2006). On logics of knowledge and belief. *Philosophical Studies, 128*(1), 169–199.

Steele, K., & Stefánsson, H. O. (2016). Decision theory. In E. N. Zalta (Ed.), *The stanford Encyclopedia of Philosophy* (Winter 2016 ed.). Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/win2016/entries/decision-theory/.

Talbert, M. (2019). Moral Responsibility. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Winter 2019 ed.). Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/win2019/entries/moral-responsibility/.

Tamminga, A. (2013). Deontic logic for strategic games. *Erkenntnis*, *78*(1), 183–200.

Thomason, R. H. (1970). Indeterminist time and truth-value gaps. *Theoria*, *36*(3), 264–281.

Thomason, R. H. (1981). Deontic logic as founded on tense logic. In *New Studies in Deontic Logic* (pp. 165–176). Springer.

Thomason, R. H. (1984). Combinations of tense and modality. In *Handbook of Philosophical Logic* (pp. 135–165). Springer.

Thompson, M. (2008). *Life and action*. Cambridge: Harvard University Press.

Tomasello, M. (2016). *A natural history of human morality*. Harvard University Press.

Troquard, N., & Balbiani, P. (2019). Propositional Dynamic Logic. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Spring 2019 ed.). Metaphysics Research Lab, Stanford University. `https://plato.stanford.edu/archives/spr2019/entries/logic-dynamic/`.

Vakarelov, D. (1992). A modal theory of arrows. arrow logics i. In *European Workshop on Logics in Artificial Intelligence* (pp. 1–24).

van Benthem, J. (2001). Games in dynamic-epistemic logic. *Bulletin of Economic Research*, *53*(4), 219–248.

van Benthem, J., & Bezhanishvili, G. (2007). Modal logics of space. *Handbook of Spatial Logics*, 217–298.

van Benthem, J., & Pacuit, E. (2014). Connecting logics of choice and change. In *Nuel Belnap on Indeterminism and Free Action* (pp. 291–314). Springer.

van Bethem, J., & Sarenac, D. (2004). The geometry of knowledge. *Travaux de Logique*, *17*, 1–31.

van de Poel, I. (2011). The relation between forward-looking and backward-looking responsibility. In *Moral Responsibility* (pp. 37–52). Springer.

van de Putte, F., Tamminga, A., & Duijf, H. (2017). Doing without nature. In *International Workshop on Logic, Rationality and Interaction* (pp. 209–223).

van der Hoek, W., & Wooldridge, M. (2002). Tractable multiagent planning for epistemic goals. In (pp. 1167–1174).

van der Hoek, W., & Wooldridge, M. (2008). Multi-agent systems. *Foundations of Artificial Intelligence*, *3*, 887–928.

van Ditmarsch, H., van der Hoek, W., Halpern, J. Y., & Kooi, B. (2015). *Handbook of epistemic logic*. College Publications.

van Sliedregt, E. (2012). *Individual criminal responsibility in international law*. OUP Oxford.

Vekony, R., Mele, A., & Rose, D. (2021). Intentional action without knowledge. *Synthese*, *199*(1), 1231–1243.

von Kutschera, F. (1986). Bewirken. *Erkenntnis*, 253–281.

von Kutschera, F. (1997). T× W completeness. *Journal of Philosophical Logic*, *26*(3), 241–250.

von Wright, G. H. (1951). Deontic logic. *Mind*, *60*(237), 1–15.

von Wright, G. H. (1963). *Norm and action: A logical enquiry*. New York, NY, USA: Routledge and Kegan Paul.

Wallace, R. J. (2020). Practical Reason. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Spring 2020 ed.). Metaphysics Research Lab, Stanford University. `https://plato.stanford.edu/archives/spr2020/entries/practical-reason/`.

Wang, Y. (2015). A logic of knowing how. In *International Workshop on Logic, Rationality and Interaction* (pp. 392–405).

Wang, Y. (2018). A logic of goal-directed knowing how. *Synthese*, *195*(10), 4419–4439.

Wansing, H. (2006a). Doxastic decisions, epistemic justification, and the logic of agency. *Philosophical Studies*, *128*(1), 201–227.

Wansing, H. (2006b). Tableaux for multi-agent deliberative-stit logic. *Advances in Modal Logic*, *6*, 503–520.

Watson, G. (1996). Two faces of responsibility. *Philosophical Topics*, *24*(2), 227–248.

Watson, G. (2001). Reasons and responsibility. *Ethics*, *111*(2), 374–394.

Weiner, B. (1995). *Judgments of responsibility: A foundation for a theory of social conduct*. Guilford Press.

Weirich, P. (2020). Causal Decision Theory. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Winter 2020 ed.). Metaphysics Research Lab, Stanford University. `https://plato.stanford.edu/archives/win2020/entries/decision-causal/`.

Willard, S. (2004). *General topology*. Courier Corporation.

Willer, M. (2012). A remark on iffy oughts. *The Journal of Philosophy*, *109*(7), 449–461.

Williamson, T. (2002). *Knowledge and its limits*. Oxford University Press.

Wölfl, S. (2002). Propositional q-logic. *Journal of Philosophical Logic*, *31*(5), 387–414.

Wooldridge, M. (2000). *Reasoning about rational agents*. Cambridge: MIT press.

Wooldridge, M., & Jennings, N. R. (1995). Intelligent agents: Theory and practice. *The Knowledge Engineering Review*, *10*(2), 115–152.

Wu, Y., Zhang, B., & Lu, J. (2011). Fuzzy logic and neuro-fuzzy systems: A systematic introduction. *International Journal OF Artificial Intelligence and Expert Systems (IJAE)*.

Xu, M. (1994). Decidability of deliberative stit theories with multiple agents. In *International Conference on Temporal Logic* (pp. 332–348).

Xu, M. (2015). Combinations of stit with ought and know. *Journal of Philosophical Logic*, *44*(6), 851–877. Retrieved from `http://dx.doi.org/10.1007/s10992-015-9365-7` doi: doi:10.1007/s10992-015-9365-7

Zanardo, A. (1985). A finite axiomatization of the set of strongly valid ockhamist formulas. *Journal of Philosophical Logic*, *14*(4), 447–468.

Zanardo, A. (1996). Branching-time logic with quantification over branches: the point of view of modal logic. *The Journal of Symbolic Logic*, *61*(1), 1–39.

Zanardo, A. (2006). Quantification over sets of possible worlds in branching-time semantics. *Studia Logica*, *82*(3), 379–400.

Samenvatting

De studie van verantwoordelijkheid is een ingewikkelde zaak. De term wordt op verschillende manieren gebruikt in verschillende vakgebieden, en het is gemakkelijk om alledaagse discussies aan te gaan over waarom iemand ergens verantwoordelijk voor moet worden gehouden. De achtergrond van deze discussies wordt meestal gevormd door sociale, juridische, morele of filosofische problemen. Een duidelijk patroon in al deze domeinen is de intentie om normen vast te stellen voor wanneer—en in welke mate—een agent verantwoordelijk moet worden gehouden voor een toestand. Dit is waar logica uitkomst biedt. De ontwikkeling van expressieve logica's, het redeneren over de beslissingen van agenten in situaties met morele gevolgen, omvat het bedenken van eenduidige representaties van componenten van gedrag die zeer relevant zijn voor de systematische toekenning van verantwoordelijkheid en voor de systematische toewijzing van schuld of lof. Om het duidelijk te stellen: expressieve syntactische en semantische kaders helpen ons om problemen gerelateerd aan verantwoordelijkheid op een methodische manier te analyseren.

Deze dissertatie bouwt een formele theorie van verantwoordelijkheid op. Het belangrijkste instrument om dit doel te bereiken is de modale logica en, in het bijzonder, een klasse van modale handelingslogica's die bekend staat als de stit-theorie. De onderliggende motivatie is om een theoretische basis te verschaffen voor symbolische technieken in de ontwerp van ethische AI. Dit werk betekent dus een bijdrage aan de formele filosofie en symbolische AI. De methodologie van het proefschrift bestaat uit de ontwikkeling van stit-theoretische modellen en talen om de wisselwerking te onderzoeken tussen de volgende componenten van verantwoordelijkheid: agency, kennis, overtuigingen, intenties en verplichtingen. Deze modellen zijn geïntegreerd in een kader dat rijk genoeg is om op logica gebaseerde karakteriseringen te bieden voor drie categorieën van verantwoordelijkheid: causale, informatieve en motivationele verantwoordelijkheid.

Het proefschrift is als volgt opgebouwd. hoofdstuk 2 bespreekt uitvoerig de stit theorie, een logica die de notie van agency in de wereld formaliseert over een indeterministische opvatting van tijd die bekend staat als branching time. Het idee is dat agenten handelen door mogelijke toekomsten te beperken tot bepaalde deelverzamelingen. Op weg naar de formalisering van informatieve verantwoordelijkheid breidt hoofdstuk 3 de stit-theorie uit met traditionele epistemische noties (kennis en geloof). Zo formaliseert het hoofdstuk belangrijke aspecten van het redeneren van agenten bij de keuze en uitvoering van handelingen. In een context van toekenning van verantwoordelijkheid en verontschuldigbaarheid breidt hoofdstuk 4 de epistemische stit-theorie uit met maatstaven voor de optimaliteit van handelingen die ten grondslag liggen aan verplichtingen. In wezen formaliseert dit hoofdstuk de wisselwerking tussen de kennis van agenten en wat zij zouden moeten doen. Op weg naar formalisering van motivationele verantwoordelijkheid voegt hoofdstuk 5 intenties en intentionele handelingen toe aan de epistemische stit theorie en redeneert over de wisselwerking tussen kennis en intentionaliteit. Tenslotte voegt hoofdstuk 6 de formalismen van de voorgaande hoofdstukken samen tot een rijke logica waarmee verschillende modi van de bovengenoemde categorieën van verantwoordelijkheid kunnen worden uitgedrukt en gemodelleerd.

Technisch gezien liggen de belangrijkste bijdragen van dit proefschrift in de axiomatiseringen van alle geïntroduceerde logica's. Met name de bewijzen van correctheid en volledigheid omvatten lange, stapsgewijze procedures die gebruik maken van nieuwe technieken.

Curriculum Vitae

Aldo Iván Ramírez Abarca was born on March 5, 1987 in Mexico City, México. He obtained a bachelor's degree in Mathematics at Universidad Nacional Autónoma de México (UNAM) in 2012, and he obtained an MSc degree in Logic at the Institute of Logic, Language, and Computation (ILLC), University of Amsterdam, in 2015. During his studies, he specialized in Topology, Analysis, Logic (specifically epistemic modal logic), and Algebra. For his master's thesis, he developed sound and complete logics of group knowledge and belief on topological models, under the tutelage of Dr. Alexandru Baltag. He started his PhD work at Utrecht University in August, 2016 on the research project titled 'Responsible Intelligent Systems,' which was led by Prof. Jan Broersen and funded by the European Research Council. As part of his doctoral research, he developed sound and complete logics of agency, practical knowledge, belief-driven choice, obligation, intentionality, and responsibility. The underlying motivation was to provide theoretical foundations for symbolic techniques in the development of ethical AI. He has taught courses on Mathematics, Logic, Artificial Intelligence, and Philosophy, and he has published in peer-reviewed conference proceedings and academic journals. His research interests include symbolic AI (responsible AI, agent-based modelling, knowledge representation, multi-agent systems), Logic (logics of action, epistemic logic, doxastic logic, deontic logic), and Formal Philosophy (epistemic game theory, decision theory).

Quaestiones Infinitae

PUBLICATIONS OF THE DEPARTMENT OF PHILOSOPHY AND RELIGIOUS STUDIES

VOLUME 21 D. van Dalen, *Torens en Fundamenten* (valedictory lecture), 1997.

VOLUME 22 J.A. Bergstra, W.J. Fokkink, W.M.T. Mennen, S.F.M. van Vlijmen, *Spoorweglogica via EURIS*, 1997.

VOLUME 23 I.M. Croese, *Simplicius on Continuous and Instantaneous Change* (dissertation), 1998.

VOLUME 24 M.J. Hollenberg, *Logic and Bisimulation* (dissertation), 1998.

VOLUME 25 C.H. Leijenhorst, *Hobbes and the Aristotelians* (dissertation), 1998.

VOLUME 26 S.F.M. van Vlijmen, *Algebraic Specification in Action* (dissertation), 1998.

VOLUME 27 M.F. Verweij, *Preventive Medicine Between Obligation and Aspiration* (dissertation), 1998.

VOLUME 28 J.A. Bergstra, S.F.M. van Vlijmen, *Theoretische Software-Engineering: kenmerken, faseringen en classificaties*, 1998.

VOLUME 29 A.G. Wouters, *Explanation Without A Cause* (dissertation), 1999.

VOLUME 30 M.M.S.K. Sie, *Responsibility, Blameworthy Action & Normative Disagreements* (dissertation), 1999.

VOLUME 31 M.S.P.R. van Atten, *Phenomenology of choice sequences* (dissertation), 1999.

VOLUME 32 V.N. Stebletsova, *Algebras, Relations and Geometries (an equational perspective)* (dissertation), 2000.

VOLUME 33 A. Visser, *Het Tekst Continuüm* (inaugural lecture), 2000.

VOLUME 34 H. Ishiguro, *Can we speak about what cannot be said?* (public lecture), 2000.

VOLUME 35 W. Haas, *Haltlosigkeit; Zwischen Sprache und Erfahrung* (dissertation), 2001.

VOLUME 36 R. Poli, *ALWIS: Ontology for knowledge engineers* (dissertation), 2001.

VOLUME 37 J. Mansfeld, *Platonische Briefschrijverij* (valedictory lecture), 2001.

VOLUME 37A E.J. Bos, *The Correspondence between Descartes and Henricus Regius* (dissertation), 2002.

VOLUME 38 M. van Otegem, *A Bibliography of the Works of Descartes (1637-1704)* (dissertation), 2002.

VOLUME 39 B.E.K.J. Goossens, *Edmund Husserl: Einleitung in die Philosophie: Vorlesungen 1922/23* (dissertation), 2003.

VOLUME 40 H.J.M. Broekhuijse, *Het einde van de sociaaldemocratie* (dissertation), 2002.

VOLUME 41 P. Ravalli, *Husserls Phänomenologie der Intersubjektivität in den Göttinger Jahren: Eine kritisch-historische Darstellung* (dissertation), 2003.

VOLUME 42 B. Almond, *The Midas Touch: Ethics, Science and our Human Future* (inaugural lecture), 2003.

VOLUME 43 M. Düwell, *Morele kennis: over de mogelijkheden van toegepaste ethiek* (inaugural lecture), 2003.

VOLUME 44 R.D.A. HENDRIKS, *Metamathematics in Coq* (dissertation), 2003.

VOLUME 45 TH. VERBEEK, E.J. BOS, J.M.M. VAN DE VEN, *The Correspondence of René Descartes: 1643*, 2003.

VOLUME 46 J.J.C. KUIPER, *Ideas and Explorations: Brouwer's Road to Intuitionism* (dissertation), 2004.

VOLUME 47 C.M. BEKKER, *Rechtvaardigheid, Onpartijdigheid, Gender en Sociale Diversiteit; Feministische filosofen over recht doen aan vrouwen en hun onderlinge verschillen* (dissertation), 2004.

VOLUME 48 A.A. LONG, *Epictetus on understanding and managing emotions* (public lecture), 2004.

VOLUME 49 J.J. JOOSTEN, *Interpretability formalized* (dissertation), 2004.

VOLUME 50 J.G. SIJMONS, *Phänomenologie und Idealismus: Analyse der Struktur und Methode der Philosophie Rudolf Steiners* (dissertation), 2005.

VOLUME 51 J.H. HOOGSTAD, *Time tracks* (dissertation), 2005.

VOLUME 52 M.A. VAN DEN HOVEN, *A Claim for Reasonable Morality* (dissertation), 2006.

VOLUME 53 C. VERMEULEN, *René Descartes, Specimina philosophiae: Introduction and Critical Edition* (dissertation), 2007.

VOLUME 54 R.G. MILLIKAN, *Learning Language without having a theory of mind* (inaugural lecture), 2007.

VOLUME 55 R.J.G. CLAASSEN, *The Market's Place in the Provision of Goods* (dissertation), 2008.

VOLUME 56 H.J.S. BRUGGINK, *Equivalence of Reductions in Higher-Order Rewriting* (dissertation), 2008.

VOLUME 57 A. KALIS, *Failures of agency* (dissertation), 2009.

VOLUME 58 S. GRAUMANN, *Assistierte Freiheit* (dissertation), 2009.

VOLUME 59 M. AALDERINK, *Philosophy, Scientific Knowledge, and Concept Formation in Geulincx and Descartes* (dissertation), 2010.

VOLUME 60 I.M. CONRADIE, *Seneca in his cultural and literary context: Selected moral letters on the body* (dissertation), 2010.

VOLUME 61 C. VAN SIJL, *Stoic Philosophy and the Exegesis of Myth* (dissertation), 2010.

VOLUME 62 J.M.I.M. LEO, *The Logical Structure of Relations* (dissertation), 2010.

VOLUME 63 M.S.A. VAN HOUTE, *Seneca's theology in its philosophical context* (dissertation), 2010.

VOLUME 64 F.A. BAKKER, *Three Studies in Epicurean Cosmology* (dissertation), 2010.

VOLUME 65 T. FOSSEN, *Political legitimacy and the pragmatic turn* (dissertation), 2011.

VOLUME 66 T. VISAK, *Killing happy animals. Explorations in utilitarian ethics.* (dissertation), 2011.

VOLUME 67 A. JOOSSE, *Why we need others: Platonic and Stoic models of friendship and self-understanding* (dissertation), 2011.

VOLUME 68 N. M. NIJSINGH, *Expanding newborn screening programmes and strengthening informed consent* (dissertation), 2012.

VOLUME 69 R. PEELS, *Believing Responsibly: Intellectual Obligations and Doxastic Excuses* (dissertation), 2012.

VOLUME 70 S. LUTZ, *Criteria of Empirical Significance* (dissertation), 2012

VOLUME 70A G.H. BOS, *Agential Self-consciousness, beyond conscious agency* (dissertation), 2013.

VOLUME 71 F.E. KALDEWAIJ, *The animal in morality: Justifying duties to animals in Kantian moral philosophy* (dissertation), 2013.

VOLUME 72 R.O. BUNING, *Henricus Reneri (1593-1639): Descartes' Quartermaster in Aristotelian Territory* (dissertation), 2013.

VOLUME 73 I.S. LÖWISCH, *Genealogy Composition in Response to Trauma: Gender and Memory in 1 Chronicles 1-9 and the Documentary Film 'My Life Part 2'* (dissertation), 2013.

VOLUME 74 A. EL KHAIRAT, *Contesting Boundaries: Satire in Contemporary Morocco* (dissertation), 2013.

VOLUME 75 A. KROM, *Not to be sneezed at. On the possibility of justifying infectious disease control by appealing to a mid-level harm principle* (dissertation), 2014.

VOLUME 76 Z. PALL, *Salafism in Lebanon: local and transnational resources* (dissertation), 2014.

VOLUME 77 D. WAHID, *Nurturing the Salafi Manhaj: A Study of Salafi Pesantrens in Contemporary Indonesia* (dissertation), 2014.

VOLUME 78 B.W.P VAN DEN BERG, *Speelruimte voor dialoog en verbeelding. Basisschoolleerlingen maken kennis met religieuze verhalen* (dissertation), 2014.

VOLUME 79 J.T. BERGHUIJS, *New Spirituality and Social Engagement* (dissertation), 2014.

VOLUME 80 A. WETTER, *Judging By Her. Reconfiguring Israel in Ruth, Esther and Judith* (dissertation), 2014.

VOLUME 81 J.M. MULDER, *Conceptual Realism. The Structure of Metaphysical Thought* (dissertation), 2014.

VOLUME 82 L.W.C. VAN LIT, *Eschatology and the World of Image in Suhrawardī and His Commentators* (dissertation), 2014.

VOLUME 83 P.L. LAMBERTZ, *Divisive matters. Aesthetic difference and authority in a Congolese spiritual movement 'from Japan'* (dissertation), 2015.

VOLUME 84 J.P. GOUDSMIT, *Intuitionistic Rules: Admissible Rules of Intermediate Logics* (dissertation), 2015.

VOLUME 85 E.T. FEIKEMA, *Still not at Ease: Corruption and Conflict of Interest in Hybrid Political Orders* (dissertation), 2015.

VOLUME 86 N. VAN MILTENBURG, *Freedom in Action* (dissertation), 2015.

VOLUME 86A P. COPPENS, *Seeing God in This World and the Otherworld: Crossing Boundaries in Sufi Commentaries on the Qurʾān* (dissertation), 2015.

VOLUME 87 D.H.J. JETHRO, *Aesthetics of Power: Heritage Formation and the Senses in Post-Apartheid South Africa* (dissertation), 2015.

VOLUME 88 C.E. HARNACKE, *From Human Nature to Moral Judgement: Reframing Debates about Disability and Enhancement* (dissertation), 2015.

VOLUME 89 X. WANG, *Human Rights and Internet Access: A Philosophical*

Investigation (dissertation), 2016.

VOLUME 90 R. van Broekhoven, *De Bewakers Bewaakt: Journalistiek en leiderschap in een gemediatiseerde democratie* (dissertation), 2016.

VOLUME 91 A. Schlatmann, *Shiʿi Muslim youth in the Netherlands: Negotiating Shiʿi fatwas and rituals in the Dutch context* (dissertation), 2016.

VOLUME 92 M.L. van Wijngaarden, *Schitterende getuigen. Nederlands luthers avondmaalsgerei als identiteitsdrager van een godsdienstige minderheid* (dissertation), 2016.

VOLUME 93 S. Coenradie, *Vicarious substitution in the literary work of Shūsaku Endō. On fools, animals, objects and doubles* (dissertation), 2016.

VOLUME 94 J. Rajaiah, *Dalit humanization. A quest based on M.M. Thomas' theology of salvation and humanization* (dissertation), 2016.

VOLUME 95 D.L.A. Ometto, *Freedom & Self-Knowledge* (dissertation), 2016.

VOLUME 96 Y. Yaldiz, *The Afterlife in Mind: Piety and Renunciatory Practice in the 2nd/8th- and early 3rd/9th-Century Books of Renunciation (Kutub al-Zuhd)* (dissertation), 2016.

VOLUME 97 M.F. Byskov, *Between experts and locals. Towards an inclusive framework for a development agenda* (dissertation), 2016.

VOLUME 98 A. Rumberg, *Transitions toward a Semantics for Real Possibility* (dissertation), 2016.

VOLUME 99 S. de Maagt, *Constructing Morality: Transcendental Arguments in Ethics* (dissertation), 2017.

VOLUME 100 S. Binder, *Total Atheism* (dissertation), 2017.

VOLUME 101 T. Giesbers, *The Wall or the Door: German Realism around 1800,* (dissertation), 2017.

VOLUME 102 P. Sperber, *Kantian Psychologism* (dissertation), 2017.

VOLUME 103 J.M. Hamer, *Agential Pluralism: A Philosophy of Fundamental Rights* (dissertation), 2017.

VOLUME 104 M. Ibrahim, *Sensational Piety: Practices of Mediation in Christ Embassy and NASFAT* (dissertation), 2017.

VOLUME 105 R.A.J. Mees, *Sustainable Action, Perspectives for Individuals, Institutions, and Humanity* (dissertation), 2017.

VOLUME 106 A.A.J. Post, *The Journey of a Taymiyyan Sufi: Sufism Through the Eyes of ʿImād al-Dīn Aḥmad al-Wāsiṭī (d. 711/1311)* (dissertation), 2017.

VOLUME 107 F.A. Fogue kuate, *Médias et coexistence entre Musulmans et Chrétiens au Nord-Cameroun: de la période coloniale Française au début du XXIème siècle* (dissertation), 2017.

VOLUME 108 J. Kroesbergen-Kamps, *Speaking of Satan in Zambia. The persuasiveness of contemporary narratives about Satanism* (dissertation), 2018.

VOLUME 109 F. Teng, *Moral Responsibilities to Future Generations. A Comparative Study on Human Rights Theory and Confucianism* (dissertation), 2018.

VOLUME 110 H.W.A. Duijf, *Let's Do It! Collective Responsibility, Joint Action, and Participation* (dissertation), 2018.

VOLUME 111 R.A. Calvert, *Pilgrims in the port. Migrant Christian communities in Rotterdam* (dissertation), 2018.

VOLUME 112 W.P.J.L. VAN SAANE, *Protestant Mission Partnerships: The Concept of Partnership in the History of the Netherlands Missionary Council in the Twentieth Century* (dissertation), 2018.

VOLUME 113 D.K. DÜRING, *Of Dragons and Owls. Rethinking Chinese and Western narratives of modernity* (dissertation), 2018.

VOLUME 114 H. ARENTSHORST, *Perspectives on freedom. Normative and political views on the preconditions of a free democratic society* (dissertation), 2018.

VOLUME 115 M.B.O.T. KLENK, *Survival of Defeat. Evolution, Moral Objectivity, and Undercutting* (dissertation), 2018.

VOLUME 116 J.H. HOEKJEN, *Pars melior nostri. The Structure of Spinoza's Intellect* (dissertation), 2018.

VOLUME 117 C.J. MUDDE, *Rouwen in de marge. De materiële rouwcultuur van de katholieke geloofsgemeenschap in vroegmodern Nederland* (dissertation), 2018.

VOLUME 118 K. GRIT, "*Christians by Faith, Pakistani by Citizenship". Negotiating Christian Identity in Pakistan* (dissertation), 2019.

VOLUME 119 J.K.G. HOPSTER, *Moral Objectivity: Origins and Foundations* (dissertation), 2019.

VOLUME 120 H. BEURMANJER, *Tango met God? Een theoretische verheldering van bibliodans als methode voor spirituele vorming* (dissertation), 2019.

VOLUME 121 M.C. GÖBEL, *Human Dignity as the Ground of Human Rights. A Study in Moral Philosophy and Legal Practice* (dissertation), 2019.

VOLUME 122 T. VAN 'T HOF, *Enigmatic Etchings. True Religion in Romeyn de Hooghe's Hieroglyphica* (dissertation), 2019.

VOLUME 123 M. DERKS, *Constructions of Homosexuality and Christian Religion in Contemporary Public Discourse in the Netherlands* (dissertation), 2019.

VOLUME 124 H. NIEBER, *Drinking the Written Qurʾan. Healing with Kombe in Zanzibar Town* (dissertation), 2020.

VOLUME 125 B.A. KAMPHORST, *Autonomy-Respectful E-Coaching Systems: Fending Off Complacency* (dissertation), 2020.

VOLUME 126 R.W. VINKESTEIJN, *Philosophical Perspectives on Galen of Pergamum: Four Case-Studies on Human Nature and the Relation Between Body and Soul* (dissertation), 2020.

VOLUME 127 L.J. JOZIASSE, *Women's faith seeking life; Lived Christologies and the transformation of gender relations in two Kenyan churches* (dissertation), 2020.

VOLUME 128 M. KRAMM, *Balancing Tradition and Development. A deliberative procedure for the evaluation of cultural traditions in development contexts* (dissertation), 2020.

VOLUME 129 N. MYLES, *Communality, Individuality and Democracy: A Defense of Personism* (dissertation), 2020.

VOLUME 130 A. OEGEMA, *Negotiating Paternal Authority and Filial Agency: Fathers and Sons in Early Rabbinic Parables* (dissertation), 2021.

VOLUME 131 A.A. GOUDRIAAN, *'Seit ein Gespräch wir sind': Language and dialogical experience in Hegel* (dissertation), 2021.

VOLUME 132 E.H. MEINEMA, *Regulating Religious Coexistence. The Intricacies of*

'Interfaith' Cooperation in Coastal Kenya (dissertation), 2021.

VOLUME 133 K.D. TIMMER, *Thresholds and limits in theories of distributive justice* (dissertation), 2021.

VOLUME 134 M.J. BLAAKMAN, *Confronting Discrimination and Unravelling the Veil of Prejudice. The epistemic conditions of responsibility for hidden prejudices* (dissertation), 2021.

VOLUME 135 A. GHAJARJAZI, *Techniques of the senses: 19th-century media and Shiism in Iran* (dissertation), 2021.

VOLUME 136 Y. AL SALMAN, *Sharing in Common: A Republican Defence of Group Ownership* (dissertation), 2022.

VOLUME 137 L.L. SCHRIJVERS, *Questioning the Conversion Paradox: Gender, Sexuality, and Belonging amongst Women Becoming Jewish, Christian, and Muslim in the Netherlands* (dissertation), 2022.

VOLUME 138 I. VAN DER GIESSEN, *Uniform Interpolation and Admissible Rules. Proof-theoretic investigations into (intuitionistic) modal logics* (dissertation), 2022.

VOLUME 139 F.H.W. BEKKERS, *De Transformatieve Dialoog. Over het oplossen van identiteitsgerelateerde morele conflicten* (dissertation), 2023.

VOLUME 140 L.M. VAN ESCH, *Sociocultural Interventions: The Performative Augmentation of the Social Imaginary* (dissertation), 2023.

VOLUME 141 A.I. RAMÍREZ ABARCA, *Logics of Responsibility* (dissertation), 2023.

Quaestiones Infinitae

θπ