Modal Logics for Rational Agents

Modale Logica's voor Rationele Actoren

(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. J.A. van Ginkel,
ingevolge het besluit van het College van Decanen
in het openbaar te verdedigen
op woensdag 19 juni 1996 des middags te 14:30 uur

door

Bernardus van Linder

geboren op 19 juni 1968
te Gemert

Promotor: Prof. dr. J.-J. Ch. Meyer
Co-promotor: Dr. W. van der Hoek
Faculteit Wiskunde en Informatica

CIP-GEGEVENS KONINKLIJKE BIBLIOTHEEK, DEN HAAG

Linder, Bernardus van

Modal logics for rational agents / Bernardus van Linder. -
Utrecht : Universiteit Utrecht, Faculteit Wiskunde en
Informatica
Proefschrift Universiteit Utrecht. - Met index, lit. opg.
- Met samenvatting in het Nederlands.
ISBN 90-393-1354-7
Trefw.: modale logica / formele talen / kunstmatige
intelligentie.

Preface

And you run and you run to catch up with the sun,
but it's sinking
And racing around to come up behind you again
The sun is the same in a relative way, but you're older.

Pink Floyd, *'Time'*.

Tracing this thesis back to its origin, I should probably start with a course on Semantics of Programming Languages, that I took in the middle of the fall semester of 1989. The hearty welcome that I received from John-Jules Meyer at that course gave me an impression of him that proved accurate over the subsequent seven years. My acquaintance with John-Jules was renewed during a course on Modal Logic, continued when writing my Masters Thesis, and finally led to me doing a Ph.D. The results captured in this thesis clearly reveal this past. The entire thesis is imbued with the things I learned in the course on Modal Logic, Chapter 4 uses elements of the course on Semantics of Programming Languages, and my Masters Thesis influenced part of Chapter 5.

The subject of this thesis is quite easily explained: it's all about the modelling of human and human-like behaviour. Models of this kind of behaviour can be used to obtain a better understanding of the way humans act, but they can also serve as specifications of entities, e.g. robots, that should display human-like behaviour. Investigating formalisms that could be used in this way has taken up the greater part of my time over the last four years. A pleasant side-effect of doing research of this kind is that one is given the opportunity to visit conferences and workshops to inform others of results that have been achieved. These travels, and in particular the meeting of different people from different cultures, provided not only a very pleasant, but also a very enriching experience.

It is a pleasure to thank all the people and institutions that contributed in one way or the other to the writing of this thesis. First and foremost I would like to thank my supervisors John-Jules Meyer and Wiebe van der Hoek. On many occasions John-Jules was far more enthusiastic than I was about the things I did, whereas Wiebe, through his sensible, Frisian, character made sure that we didn't get too enthusiastic. Not only did I learn a lot from working with John-Jules and Wiebe, it was also good fun. Furthermore my thanks go out to my co-authors, who are, in addition to John-Jules and Wiebe, Peter

Bruza, Frank Dignum and Theo Huibers, for a very pleasant and fruitful cooperation. I want to thank the members of my reading committee who are, in alphabetical order, Prof. Dr Jan van Leeuwen, Prof. Dr Raymond Reiter, Dr Maarten de Rijke, Dr Pierre-Yves Schobbens and Prof. Dr Jan Treur, for their careful proofreading of this thesis. Thanks are also due to the members of the Theoretical Computer Science section of the Vrije Universiteit and the Department of Computer Science of Utrecht University for providing both a stimulating and a pleasant working environment. Visiting conferences can be rather expensive. I am therefore very grateful to the Departments of Computer Science of the Vrije Universiteit and of Utrecht University, to the Faculty of Information Technology at the Queensland University of Technology, to the Esprit II BRA 'Drums II' and Esprit III BRWG 'ModelAge' research projects, to NWO and to the Shell Travel Fund for providing (partial) coverage of the expenses associated with attending and speaking at conferences and workshops.

On a personal level I would like to thank some very good friends, who were always there for life outside of Academia. In particular I thank the *Gorinchem Gang*, Bas, Leonard, Ronald and Margriet *cum suii*, and the *Nijmegen Nerds*, Ton, Thomas and Christ, for many a pleasant gathering. Special thanks go out to the *three T's*, Thomas, Ton, and Theo. Thomas not only endured my moods during the week, but was also still able to remain a friend both in and outside University. I've never met anybody as cheerful and attentive as Thomas, which makes him the perfect office mate and friend. With Ton I exchanged over 3000 email messages, either with or without E. Weber, and furthermore organised 'Kronenberg-Puiflijk-Ulm Treffen', which provided lousy games of chess and very deep conversations. Theo is not only a fine colleague, an enthusiastic co-author and a good friend, but in addition is the most hospitable guy I have ever met. Theo and Pio showed the true meaning of 'Brabantse gastvrijheid'. Last, but definitely not least, I would like to thank my sister, my father, and my mother for their constant encouragement and interest in my activities. My sister kept me down to earth by constantly telling me that computer scientists are nerds: there is no escaping from that. My father taught me that there is nothing wrong with having ambitions and trying to fulfil them. My mother reminded me that doing your best is what matters most, and that achieving specific ambitions is not everything.

It was a pleasure *Bernd van Linder, Utrecht, April 1996.*

Contents

Chapter 1

Introduction

The idea of building this kind of functionality into a computer until recently was a dream so far out of reach that the concept was not taken seriously. This is changing rapidly. Enough people now believe that such 'interface agents' are buildable. For this reason, this backwater interest in intelligent agents has become the most fashionable topic of research in human-computer interface design. It has become obvious that people want to delegate more functions and prefer to directly manipulate computers less.

Nicholas Negroponte, *'Being Digital'*.

In this chapter we lay the foundations of this thesis. We present our personal view on agents and agency, and indicate the importance of formal semantics when dealing with (implementations of) rational agents. It is argued that modal logic provides a good formal tool to model agency. Thereafter the scope of this thesis is delineated, and some preliminary concepts are defined. The chapter is concluded with an overview of what is still to come.

1.1 What's an agent, anyway?

The last ten years have witnessed an intense flowering of interest in agents, both on a theoretical and on a practical level. The ACM devoted a special issue of its 'Communications' to intelligent agents [21], and Scientific American ranked intelligent software agents among the key technologies for the 21st century [89]. Also various conferences and workshops were initiated that specifically address agents, their theories, languages, architectures and applications [32, 58, 79, 130, 131]. Consequently, terms like agent-based

computing, agent-based software engineering and agent-oriented programming have become widely used in research on AI. Despite its wide use, there is no agreement on what the term 'agent' means. Riecken remarks that 'at best, there appears to be a rich set of emerging views' and that 'the terminology is a bit messy' [114]. Existing definitions range from 'any entity whose state is viewed as consisting of mental objects ' [121] and 'any entity having the ability to pursue goals, to control its action and in some way to communicate with other agents' [98], to 'autonomous objects with the capacity to learn, memorize and communicate' [33] and 'systems whose behavior is neither casual nor strictly causal, but teleonomic, goal-oriented toward a certain state of the world' [16]. Other authors, and truly not the least, use the term 'robot' instead of agent [78], or take the commonsense definition of agents for granted [109]. In practical applications agents are 'personal assistent[s] who [are] collaborating with the user in the same work environment' [88], or 'computer programs that simulate a human relationship, by doing something that another person could otherwise do for you' [119]; in applications dealing with Virtual Reality so-called believable agents are seen as 'intelligent, emotional, behaving creatures [that] must respond to the user's rich variety of human behavior in believable ways' [8].

The informal description of an agent in its most primitive form, which we distil from the definitions given above and which the reader is kindly advised to keep at the back of his/her mind throughout reading this thesis, is that of an entity which has the possibility to execute certain *actions*, and is in the possession of certain *information*, which allows it to *reason* about its own and other agents' actions and the *motives* underlying these acts. In general, agents will not be this primitive, but will also be *rational*, both in their behaviour and with respect to the information that they posses, *autonomous*, in that they can to a certain extent make their own decisions on what to do next, *social* in their behaviour with respect to other agents, and in addition have the capacity to *acquire new information* and *adapt their behaviour* in the light of this new information. As we see it, the concrete agents in which all of these aspects are embodied are either human or highly sophisticated artificial agents. The formalisation of this kind of agents is the subject of this thesis.

1.2 The importance of formal methods

Currently several applications of agent-technology are in use. Among those listed by Wooldridge & Jennings [129] are air-traffic control systems, spacecraft control, telecommunications network management and particle acceleration control. Furthermore, interface agents are used that for instance take care of email administration, as well as information agents that deal with information management and retrieval. In all probability, these implemented agents will be rather complex. In addition, life-critical im-

plementations like air-traffic control systems and spacecraft control systems need to be highly reliable. We agree with Butler & Finelli [12, 13, 128] that the only possible way to guarantee reliability of such complex systems is by using formal methods in their development. This guarantee of reliability can never be given by just performing tests on the system. Besides these general reasons for using formal techniques in any branch of AI and computer science, there is another reason when dealing with agents. These agents will in general be equipped with features representing commonsense concepts as knowledge, belief and ability. Since most people do have their own conception of these concepts, it is very important to unambiguously establish what is meant by these concepts when ascribed to some specific implemented agent. Formal specifications allow for such an unambiguous definition.

Very nice examples of the use and usability of formal methods when dealing with rational agents are given by Jones [64] and by Krogh [74]. Jones showed, by rigorously formalising some assumptions that were (implicitly) underlying a certain incarnation of agent-oriented programming [120], the irrationality of the, supposedly rational, agents that had been implemented. More in detail, Jones proved that, under the given assumptions, any agent which cannot do something, believes that it can. In turn, Krogh formally showed that another incarnation of agent-oriented programming [121] suffered from the problem that different agents were not allowed to have conflicting obligations, a property which is, though possibly desirable, intuitively highly unrealistic. And even though it is quite possible that this irrational and unrealistic behaviour could also have been noticed by testing the implemented agents, in both cases it actually was the use of formal methods that did so.

1.2.1 Modal logics for agency

The formal tool that we propose to model agency is *modal logic* [19, 60, 61]. Originally being the logic of Leibnizian necessity and possibility, modal logics have been used extensively in formalising intensional notions in analytical philosophy [35]. It is perhaps even not too bold to say that modal logic has by now become *the* standard paradigm for the formal analysis of this kind of notions in a philosophical context. Also in theoretical computer science modal logics are used to analyse, specify and verify all kinds of computer programs and automated systems [30, 70].

Using modal logics offers a number of advantages. Firstly, using an intensional logic like modal logic allows one to come up with an intuitively acceptable formalisation of intensional notions with much less effort than it would take to do something similar using full-fledged first-order logic. Secondly, the reducibility of modal logic to (fragments of) first-order logic ensures that methods and techniques developed for first-order logic are still applicable to modal logic. Lastly, using possible worlds models as originally proposed

by Kripke [72], provides for a uniform, clear, intelligible, and intuitively acceptable means to give mathematical meaning to a variety of modal operators.

1.3 What this thesis is about

In this thesis we study the usability of modal logic as a tool to formalise rational agents. To this end we combine and extend various modal logics , viz. *epistemic* logic, the logic of knowledge, *doxastic* logic, the logic of belief, and *dynamic* logic, the logic of action. More in particular, we extend doxastic logic to account for different degrees of credibility and reliability of beliefs, and we propose a new paradigm for dynamic logic that generalises the standard one. In addition, we propose a formalisation of ability that allows one to formally distinguish the notions of opportunity and ability, which are, according to insights gained in the research on analytical philosophy, indeed essentially different. The logics used in defining the formal system are subject of research in mathematics, analytical philosophy and computer science, while the algorithmic character of the solutions that solve the problems encountered in extending these logics clearly reveals their roots in computer science. As such, the research captured in this thesis is situated at the crossroads of various disciplines relevant to AI.

Our main aim in defining the formal system is to end up with a language that is highly expressive yet intelligible and concise, with a clear and well-defined semantics, which allows one to model very rich theories of agency. This system is to be used by theorists to gain a better understanding and insight into the nature of agents and agency. The resulting framework is a very flexible one, which contains, in the form as it is presented in this thesis, many personal, and possibly arguable, choices. Due to its enormous flexibility one is however not forced to maintain these choices when applying the framework; it is easily tuned to anyone's personal preferences. As such the contribution of this thesis is twofold: on the one hand we present a very flexible formal system that can be used to model all kinds of aspects of agency, on the other hand we propose a personal instantiation of this system which clearly shows its usability.

1.3.1 What this thesis is not about

Throughout this thesis logic is treated as a useful formalisation and specification tool, with no intrinsic interest of its own, which makes the research captured in this thesis to belong to the field of *applied logic* rather than *pure logic*. As such we do not extensively deal with topics in pure logic like decidability, complexity, compactness and so on.

At the other side of the spectrum we do also not consider genuine practical issues. Although we give examples of possible specifications at various places throughout this thesis, we do not put our formalism to the test of specifying existing (software) agents.

The formal specification of this kind of agents should constitute the major part of future research.

1.4 Guide to the reader

For the greater part this thesis is meant to be self-contained. Basically we only assume the reader to be familiar with both (classical) propositional logic and (classical) first-order logic; for those who are not yet so, good introductions are given by Gamut [36, 37] and Leblanc & Wisdom [76]. Furthermore, some familiarity with set theory is requested, but this will not exceed the material given in standard textbooks. Knowledge of modal logic is not a prerequisite, though on occasion it might come in handy.

Although this thesis is all about formalising notions, the original aim upon writing it was to make it readable even for those with an aversion to formulae and formalisms. Just by reading the text while skipping all formulae one should be able to at least get a very good impression of what is going on. And armed with such an impression, on second reading the formulae might turn out not to be too intimidating after all.

This thesis contains a considerable number of propositions, theorems, lemmas and corollaries. Although all of these were meticulously proved, not all of these proofs are actually given here. Some proofs are omitted for being too trivial to bother the reader with, other omitted proofs are completely analogous to proofs that are given and are therefore left out, and yet other proofs are too elaborate and space-consuming to be spelled out[1]. Even though some proofs are left out, the reader should at all times be aware that all claims made in this thesis are correct to the best of our knowledge and have been proved to be so to the best of our capacities.

1.4.1 Formal preliminaries

As mentioned above, we assume the reader to be familiar with classical propositional and first-order logic. More in particular, we assume familiarity with both the syntax and the semantics, as well as with the proof systems of these logics. All concepts additional to these are introduced with ample explanation and as much care as possible. The only formal concepts that are used in this thesis without explanation — because this would either cause too much of an interruption or fall outside the scope of this thesis — are the following ones. The symbol '$\triangleq$' denotes *definitional equality* and is used to define new expressions in terms of previously existing ones. The symbol '$\wp$' is used to denote the *power set* of a given set; for V an arbitrary set, $\wp(\mathrm{V})$ denotes the power set of V. Analogously, $\mathbf{.}$ is used to denote the *lift* of a given set. For an arbitrary set V, the set

[1] In this case, however, these proofs are given elsewhere, and exact pointers to their locations are provided.

$\mathbf{.}(\mathrm{V})$, which is in general denoted by $\mathrm{V}^{\cdot}$, is defined to be $\mathrm{V} \cup \{\emptyset\}$. We furthermore assume the *truth-values* $\mathbf{1}$ and $\mathbf{0}$, as well as the set bool containing these values, to be given. The function Cn yields the set of consequences in classical propositional logic of a given set of propositional formulae, i.e. if Φ is a set of propositional formulae then $\mathrm{Cn}(\Phi)$ consists of all propositional formulae that can be derived from Φ in classical propositional logic.

In reasoning at a meta-level about the properties of the formal systems presented in this thesis, we use the symbol $\Rightarrow$ to denote 'only if', $\Leftarrow$ to denote 'if', $\Leftrightarrow$ to denote 'if and only if', and $\&$ to denote 'and'. Brackets will usually be dropped as much as possible without causing confusion. To this end we will assume that any element from $\{\rightarrow, \leftarrow, \leftrightarrow\}$ has a lower priority than one from $\{\wedge, \vee\}$, and that the same holds for $\{\Rightarrow, \Leftarrow, \Leftrightarrow\}$ with respect to $\{\&$,or$\}$.

1.4.2 Outline

This thesis contains, in addition to this prefatory one, six more chapters. Except for the next one, each of these chapters contains the definitions that constitute one or more formal systems, which focus on a particular aspect of agency. In the following chapter, which is called *Modelling rational agents*, the philosophical foundations that we consider fundamental to agency, viz. knowledge, ability, opportunity and result, are laid out. Furthermore, a formalisation of these and other notions constituting the core of all systems defined in subsequent chapters, is presented.

In the third chapter, *A logic of capabilities*, the first two of the formal systems built on the core defined in Chapter 2 are presented. These systems, belonging to the so-called KARO-architecture, share a common language and class of models, but differ in their interpretation of certain formulae. Using the primitive concepts of these systems we define the notion of practical possibility, which is not only philosophically interesting, but also from the point of view of AI, since it can be used to formalise part of the planning of agents. For validity in both systems a sound and complete axiomatisation is presented, the most striking feature of which is the use of infinitary proof rules, i.e. rules with an infinite number of premises.

The fourth chapter is called *Unravelling nondeterminism*. In this chapter we investigate, and present solutions to, the problems associated with nondeterminism of actions. We consider two kinds of nondeterminism, viz. internal nondeterminism in which the (nondeterministic) choice is up to the agent, and external nondeterminism where some unspecified external environment decides which of two nondeterministically composed actions is to be performed. The presence of ability as a primitive notion in the core of all of our formal systems renders most of the standard approaches towards the formalisation of nondeterminism useless; only for one of the systems defined in Chapter 3 in combination with one particular kind of nondeterminism, a more or less standard approach

is applicable. To allow for other combinations of the systems of Chapter 3 and various kinds of nondeterminism, we present some pragmatic, algorithmic definitions that indeed fulfil our purposes.

In the fifth chapter, *Intelligent information agents*, we focus on the formalisation of both the statics and dynamics of information from an agent-oriented perspective. Concerning the statics, we present a formalisation of belief that allows for a classification according to credibility. As to the dynamics, we consider various so-called informative actions that model different ways of information acquisition open to an agent. More in particular, we formalise observations, communication and the jumping to conclusions that constitutes default reasoning, as actions to be executed by the agents. Different ways of acquiring information result in different degrees of credibility attached to the information, which is formalised using the credibility classification of the agents' beliefs. To interpret informative actions, and in particular to ensure that execution of such an action changes the beliefs of an agent in a correct way, we propose a generalisation of the standard paradigm of dynamic logic. Employing this more general paradigm we indeed succeed in coming up with an adequate formalisation.

The sixth chapter, which is called *How to motivate your agents*, is devoted to a formalisation of motivational attitudes, the attitudes that explain why agents act the way they do. The most remarkable aspect of this formalisation is the wide range of attitudes that are considered. Not only do we deal both with attitudes that range over propositions, viz. wishes and goals, and with attitudes that range over actions, viz. commitments, but we also consider both the statics and dynamics of these attitudes. Basically, the idea is that an agent's wishes constitute its primitive motivational attitudes, and that its goals are selected among its unfulfilled yet implementable wishes. On the basis of its knowledge of its goals and practical possibilities, an agent may make certain commitments. These commitments are in general persistent, but as soon as one of its commitments is worn out an agent has the possibility to undo it. The act of selecting wishes and those of committing and undoing commitments are modelled according to the new paradigm for dynamic logic as introduced in Chapter 5.

The thesis is concluded with a chapter in which we reflect on the picture of agents as emerging from this thesis. To show how the formal machinery presented in the Chapters 2 through 6 could be applied in practice, we sketchily specify an artificial information agent. Finally, we recapitulate the main achievements and conclusions, and indicate directions for future research.

In the light of the interdependencies of the chapters, one is strongly advised to read Chapter 2 before reading any other chapter. Furthermore, reading Chapter 3 should precede reading Chapter 4, and before reading Chapter 6 both Chapter 4 and Chapter 5 should be read. Finally, by its very nature, Chapter 7 is best read in the end.

Chapter 2

Modelling rational agents

When I found so astonishing a power placed within my hands, I hesitated a long time concerning the manner in which I should employ it. Although I possessed the capacity of bestowing animation, yet to prepare a frame for the reception of it, with all its intricacies of fibres, muscles, and veins, still remained a work of inconceivable difficulty and labour. I doubted at first whether I should attempt the creation of a being like myself, or one of simpler organisation; but my imagination was too much exalted by my first success to permit me to doubt of my ability to give life to an animal as complex and wonderful as man.

Mary Shelley, '*Frankenstein*'.

As mentioned in Chapter 1, an agent is an entity which has the possibility to perform certain actions and is in the possession of certain information, which it may use to reason about its acts. We start off this chapter by explaining the notions that we consider fundamental to agency, viz. knowledge and actions. For this explanation we draw upon insights acquired in analytic philosophy, without having the pretension that our account provides *the* philosophical theory of knowledge and action. Its mere purpose is to provide the reader with an intuitive grasp for the notions that we will formalise. The formalisation of the core notions that will form the foundation of all systems presented in this thesis constitutes the rest of this chapter. We use the term 'system' rather loosely to describe the complex of language, models, interpretations and occasionally proof theory. The elements of a system can roughly be classified as belonging to the syntax or to the semantics. Whereas the syntax defines the well-formed formulae that constitute the language, the semantics is concerned with giving mathematical meaning to these formulae. In addition to the core syntax and semantics we present a uniform account

of a variety of definitions that are used at various occasions throughout the rest of this thesis. We conclude this chapter with a brief summary and some references to the — mainly philosophical — literature.

2.1 Knowledge, action and events from a philosophical perspective

The most fundamental informational attitude that we equip our agents with is termed *knowledge*. Knowledge is one of those commonsense terms that are intuitively understood by virtually everybody, yet extremely hard to characterise, describe and analyse[1]. Our use of the term knowledge complies with the common one in AI and computer science [45, 96], i.e. knowledge is veridical information on which the agent has both positive and negative introspection. Veridicality implies that only true formulae are known by agents, positive introspection states that agents know that they know something whenever they know it, and negative introspection states that agents know that they do not know something as soon as they do not know it. On occasion, viz. in Chapter 5, we will think of the agents' knowledge as being *a priori*, i.e. belonging to pure reason. However, for our purposes it is in general irrelevant whether one thinks of knowledge as being *a priori*, or *a posteriori*, i.e. due to practical experience[2].

To explain the concept of action, we first have to spend some words on the ontology of states of affairs that we presuppose. A state of affairs is assumed to be a description of the world. This description may both be actual, i.e. pertaining to the real world, or hypothetical, for example in the case of an agent which considers various descriptions of the world possible on the basis of its knowledge. Actions are now considered to be descriptions of causal processes, which upon execution by an agent may turn one state of affairs into another one. As such our intuitive idea of actions corresponds to what Von Wright calls the *generic* view on actions [133]. An *event* consists of the performance of a particular action by a particular agent, and is as such related to Von Wright's *individual* view on actions [133]. Given the ontology of actions and events as somehow causing transitions between states of affairs, we deem two aspects of these notions to be crucial: when is it possible for an agent to perform an action, and what are the effects of the event consisting of the performance by a particular agent of a particular action in a particular state of affairs? To investigate these questions we focus on three aspects of actions and events that are in our opinion essential, viz. *result*, *opportunity* and *ability*. Slightly simplifying ideas of Von Wright [132], we consider any aspect of the state of affairs brought about by the occurrence of an event in some state of affairs to be among the result of that particular event in that particular state of affairs. In adopting

[1] Another example of such a term is *intelligence*, making *knowledge* representation in artificial *intelligence* to be among the most fascinating, interesting and inherently difficult areas of research.

[2] Used in this way, the terms *a priori* and *a posteriori* are due to Kant [66].

this description of results we abstract from all kinds of aspects of results that would probably have to be dealt with in order to come up with an account that is completely acceptable from a philosophical point of view, like for instance the question whether all changes in a state of affairs have to be ascribed to the occurrence of some event, thereby excluding the possibility of external factors influencing these changes. However, it is not our aim to provide a thorough and complete theory of results, but instead combine results with other notions that are important for agency. From this point of view it seems that our definition of results is sufficiently adequate to investigate the effects of actions, and, given the complexity already associated with this simple definition, it does not make much sense to pursue even more complex ones.

Just as that of the result of events, the notions of ability and opportunity are among the most discussed and investigated in analytical philosophy. Ability plays an important part in various philosophical theories, as for instance the theory of free will and determinism, the theory of refraining and seeing-to-it, and deontic theories. Following Kenny [68], we consider ability to be the complex of physical, mental and moral capacities, internal to an agent, and being a positive explanatory factor in accounting for the agent performing an action. Opportunity on the other hand is best described as circumstantial possibility, i.e. possible by virtue of the circumstances. The opportunity to perform some action is external to the agent and is often no more than the absence of circumstances that would prevent or interfere with the performance. Although essentially different, abilities and opportunities are interconnected in that abilities can be exercised only when opportunities for their exercise present themselves, and opportunities can be taken only by those who have the appropriate abilities. From this point of view it is important to remark that abilities are understood to be *reliable* (cf. [10]), i.e. having the ability to perform a certain action suffices to take the opportunity to perform the action every time it presents itself. The combination of ability and opportunity determines whether or not an agent has the (practical) possibility to perform an action.

2.2 Knowledge, action and events from a formal perspective

In formalising the notions discussed in the previous section, we will concentrate on some rather precisely described aspects, thereby totally ignoring others. In our formalisation of knowledge we will for instance leave — nonetheless important — aspects as logical omniscience and bounded rationality out of consideration, and instead assume our agents to be perfect logical reasoners. From our point of view, knowledge is mainly important as formalising the information of agents on their abilities, opportunities and the results of their actions, without too much intrinsic interest of its own. Therefore, the notion of knowledge as described above suffices for our purposes. Concerning the (regular) actions of agents, we will in general not try to formalise what it means exactly to have the ability

or opportunity to perform some action, but instead focus on the compositional behaviour of actions, i.e. how does the ability or opportunity to perform some composite action relate to the ability or opportunity for the components of this actions. In addition, we introduce new, special actions that formalise commonsense acts as observing and communicating, and single out the constituents of the opportunity, ability and result of these actions. For the reasons already given in Chapter 1, we propose the use of a propositional multi-modal language to formalise all of these aspects. In contrast with most philosophical accounts, but firmly in the tradition of theoretical computer science, this language is an *exogenous* one, i.e. actions are represented explicitly. Although it is certainly possible to come up with accounts of action without representing actions (see for instance [65, 104, 105, 118]), we are convinced that many problems that plague these endogenous formalisations can be avoided in exogenous ones[3]. The core of each class of actions considered in this thesis is built up from a set of atomic actions using a variety of constructors. In our perception, one of the defining characteristics of atomic actions is their determinism, i.e. there is at most one state of affairs brought about by execution of an atomic action in some (other) state of affairs. The constructors that we propose to build more complex actions out of the atomic ones, deviate somewhat from the standard actions from dynamic logic [43, 46], but are both well-known from high-level programming languages and somewhat closer to philosophical views on actions than the standard constructors.

The multi-modal language that forms the core of all formalisations that are subsequently investigated in this thesis, contains modal operators to represent the knowledge of agents as well as to represent the result and opportunity of events. The ability of agents is formalised by a factually non-modal operator. Following the representation of Hintikka [48] we use the operator $\mathbf{K}\_$ to refer to the agents' knowledge: $\mathbf{K}_i\varphi$ denotes the fact that agent i knows φ to hold. To formalise results and opportunities we borrow constructs from dynamic logic: $\langle \mathrm{do}_i(\alpha)\rangle\varphi$ denotes that agent i has the opportunity to perform the action α in such a way that φ will result from this performance. The abilities of agents are formalised through the $\mathbf{A}\_$ operator: $\mathbf{A}_i\alpha$ states that agent i has the ability to perform the action α.

In subsequent chapters we define various formal systems. To make explicit that the system of this chapter is but the core of all the genuine, fully-fledged systems presented subsequently, we attach a generic variable X as a superscript to some of the notions defined below. By either omitting X or instantiating in a suitable manner, we immediately obtain the appropriate incarnation of these core notions.

The language $\mathrm{L}_0(\Pi)$, which is the language of classical propositional logic, does not

[3]Similar observations underly the (partial) shift from endogenous logics of agency [109] to exogenous ones [107] in the work of Rao & Georgeff, and in Meyer's (exogenous) dynamic deontic logic [92], which avoids much of the problems that plague (endogenous) deontic logics [3].

depend on the system under consideration, and is therefore defined once and for all in Definitions 2.1 and 2.3.

2.1. Definition. The propositional language $\mathrm{L}_0(\Pi)$ is founded on a set Π of propositional symbols. The alphabet of $\mathrm{L}_0(\Pi)$ contains the well-known connectives $\neg$ and $\wedge$.

2.2. Definition. The language $\mathrm{L}^{\mathrm{X}}(\Pi, \mathrm{A}, \mathrm{At})$ is founded on three denumerable, non-empty sets, each of which is disjoint of the others: Π is the set of propositional variables, $\mathrm{A} \subseteq \mathbb{N}$ is the set of agents, and At is the set of atomic actions. The core alphabet contains the well-known connectives $\neg$ and $\wedge$, the epistemic operator $\mathbf{K}\_\_$, the dynamic operator $\langle \mathrm{do}\_(\_)\rangle\_$, the ability operator $\mathbf{A}\_\_$, the action constructors `confirm` $\_$ (confirmations), $\_;\_$ (sequential composition), `if_then_else_fi` (conditional composition) and `while_do_od` (repetitive composition).

2.3. Definition. The language $\mathrm{L}_0(\Pi)$ is the smallest superset of Π such that

- if $f \in \mathrm{L}_0(\Pi), g \in \mathrm{L}_0(\Pi)$ then $\neg f \in \mathrm{L}_0(\Pi)$, $f \wedge g \in \mathrm{L}_0(\Pi)$

The language $\mathrm{L}^{\mathrm{X}}(\Pi, \mathrm{A}, \mathrm{At})$ is a superset of Π such that

- if $\varphi \in \mathrm{L}^{\mathrm{X}}(\Pi, \mathrm{A}, \mathrm{At})$ and $\psi \in \mathrm{L}^{\mathrm{X}}(\Pi, \mathrm{A}, \mathrm{At})$ then $\neg\varphi \in \mathrm{L}^{\mathrm{X}}(\Pi, \mathrm{A}, \mathrm{At})$ and $\varphi \wedge \psi \in \mathrm{L}^{\mathrm{X}}(\Pi, \mathrm{A}, \mathrm{At})$
- if $\varphi \in \mathrm{L}^{\mathrm{X}}(\Pi, \mathrm{A}, \mathrm{At})$, $i \in \mathrm{A}$, $\alpha \in \mathrm{Ac}^{\mathrm{X}}(\mathrm{At})$ then $\mathbf{K}_i\varphi \in \mathrm{L}^{\mathrm{X}}(\Pi, \mathrm{A}, \mathrm{At})$, $\langle \mathrm{do}_i(\alpha)\rangle\varphi \in \mathrm{L}^{\mathrm{X}}(\Pi, \mathrm{A}, \mathrm{At})$ and $\mathbf{A}_i\alpha \in \mathrm{L}^{\mathrm{X}}(\Pi, \mathrm{A}, \mathrm{At})$

where the class $\mathrm{Ac}^{\mathrm{X}}(\mathrm{At})$ of actions is a superset of At such that

- if $\varphi \in \mathrm{L}^{\mathrm{X}}(\Pi, \mathrm{A}, \mathrm{At})$ then $\mathtt{confirm}\,\varphi \in \mathrm{Ac}^{\mathrm{X}}(\mathrm{At})$
- if $\alpha_1 \in \mathrm{Ac}^{\mathrm{X}}(\mathrm{At}), \alpha_2 \in \mathrm{Ac}^{\mathrm{X}}(\mathrm{At})$ then $\alpha_1; \alpha_2 \in \mathrm{Ac}^{\mathrm{X}}(\mathrm{At})$
- if $\varphi \in \mathrm{L}^{\mathrm{X}}(\Pi, \mathrm{A}, \mathrm{At})$, $\alpha_1 \in \mathrm{Ac}^{\mathrm{X}}(\mathrm{At}), \alpha_2 \in \mathrm{Ac}^{\mathrm{X}}(\mathrm{At})$ then $\mathtt{if}\,\varphi\,\mathtt{then}\,\alpha_1\,\mathtt{else}\,\alpha_2\,\mathtt{fi} \in \mathrm{Ac}^{\mathrm{X}}(\mathrm{At})$
- if $\varphi \in \mathrm{L}^{\mathrm{X}}(\Pi, \mathrm{A}, \mathrm{At})$, $\alpha \in \mathrm{Ac}^{\mathrm{X}}(\mathrm{At})$ then $\mathtt{while}\,\varphi\,\mathtt{do}\,\alpha\,\mathtt{od} \in \mathrm{Ac}^{\mathrm{X}}(\mathrm{At})$

Additional constructs are introduced by definitional abbreviation:

$\varphi_1 \vee \varphi_2$	$\triangleq$	$\neg(\neg\varphi_1 \wedge \neg\varphi_2)$
$\varphi_1 \to \varphi_2$	$\triangleq$	$\neg\varphi_1 \vee \varphi_2$
$\varphi_1 \leftrightarrow \varphi_2$	$\triangleq$	$(\varphi_1 \to \varphi_2) \wedge (\varphi_2 \to \varphi_1)$
$\top$	$\triangleq$	$p \vee \neg p$ for arbitrary $p \in \Pi$
$\bot$	$\triangleq$	$\neg\top$
$\mathbf{M}_i\varphi$	$\triangleq$	$\neg\mathbf{K}_i\neg\varphi$
$[\mathrm{do}_i(\alpha)]\varphi$	$\triangleq$	$\neg\langle \mathrm{do}_i(\alpha)\rangle\neg\varphi$
`skip`	$\triangleq$	$\mathtt{confirm}\,\top$
`fail`	$\triangleq$	$\mathtt{confirm}\,\bot$
α^0	$\triangleq$	`skip`
α^{n+1}	$\triangleq$	$\alpha; \alpha^n$

The following letters, possibly marked, are used as typical elements:

- p, q, r for the elements of Π
- i, j for the elements of A
- a, b, c for the elements of At
- f, g, h for the elements of $\mathrm{L}_0(\Pi)$
- φ, ψ, ρ for the elements of $\mathrm{L}^{\mathrm{X}}(\Pi, \mathrm{A}, \mathrm{At})$
- α, β, γ for the elements of $\mathrm{Ac}^{\mathrm{X}}(\mathrm{At})$

Whenever the sets $\Pi, \mathrm{A}, \mathrm{At}$ are understood, which we assume to be the case unless explicitly stated otherwise, we write L_0, L^{X} and Ac^{X} rather than $\mathrm{L}_0(\Pi)$, $\mathrm{L}^{\mathrm{X}}(\Pi, \mathrm{A}, \mathrm{At})$ and $\mathrm{Ac}^{\mathrm{X}}(\mathrm{At})$.

2.4. Definition. The clauses given above in the definition of the language $\mathrm{L}^{\mathrm{X}}(\Pi, \mathrm{A}, \mathrm{At})$ and the class Ac^{X} of actions will be used in defining each of the languages considered in the sequel of this thesis, and are therefore termed core clauses.

The elements of Π and At denote (uninterpreted) propositional symbols — or atomic formulae — and atomic actions, respectively. The elements of A are natural numbers representing the agents under consideration. The constructs $\neg, \wedge, \vee, \rightarrow, \leftrightarrow$ and $\top, \bot$, denoting negation, conjunction, disjunction, implication, equivalence and the canonical tautology and contradiction, respectively, are all well-known from classical (propositional) logic. The intuitive interpretation of formulae $\mathbf{K}_i\varphi$, $\langle \mathrm{do}_i(\alpha) \rangle \varphi$ and $\mathbf{A}_i\alpha$ is discussed above. The formula $\mathbf{M}_i\varphi$ is the dual of $\mathbf{K}_i\varphi$ and represents the epistemic possibility of φ for agent i, i.e. on the basis of its knowledge, i considers φ to be possible. The formula $[\mathrm{do}_i(\alpha)]\varphi$ is the dual of $\langle \mathrm{do}_i(\alpha) \rangle \varphi$; this formula is noncommittal about the opportunity of agent i to perform the action α but states that if the opportunity to do α is present, then φ would be among the results of $\mathrm{do}_i(\alpha)$. The action constructors presented in Definition 2.3 constitute the class of so-called *strict programs* (cf. [44, 46]). Their intuitive interpretation is as follows:

`confirm` φ	verify φ
α_1; α_2	α_1 followed by α_2
`if` φ `then` α_1 `else` α_2 `fi`	α_1 if φ holds and α_2 otherwise
`while` φ `do` α `od`	α as long as φ holds

The action `skip` represents the void action, and `fail` denotes the abort action. The action α^n consists of sequentially doing α n times.

As is shown in the following example, solely using the constructs present in the core language L^{X} already allows one to formalise fairly elaborate, pseudo-realistic situations.

2.5. Example. Consider the following action: 'If some software agent i knows that the user u that it is assisting will feel better after being helped by i and i knows that it is

able to do so, then it will help its user and otherwise it will do nothing'. Actions like these could well be thought of as occurring in specifications of software agents. Using constructs from L^{X} it is possible to formalise the situation that if u is feeling well before i performs the aforementioned action, then u will still feel well or u will feel even better after i has performed the action. Denoting 'u better', 'u well' and 'help' by b, w and h, respectively, and assuming that $\{w, b\} \subseteq \Pi$, $i \in \mathrm{A}$ and $h \in \mathrm{At}$, this formalisation amounts to: $w \to [\mathrm{do}_i(\mathtt{if}\ (\mathbf{K}_i[\mathrm{do}_i(h)]b \wedge \mathbf{K}_i\mathbf{A}_i h)\ \mathtt{then}\ h\ \mathtt{else}\ \mathtt{skip}\ \mathtt{fi})](w \vee b)$.

The vast majority of all interpretations proposed for modal languages is based on the use of Kripke-style possible worlds models. In general, these models consist of a non-empty set of possible worlds, which may be thought of as the formal counterparts of the states of affairs already discussed in our description of action, a valuation on propositional symbols indicating the truth value of atomic propositions in possible worlds, and various accessibility relations between these worlds, of which the interpretation depends on the actual area of application: for the standard modal logic of possibility and necessity, this relation is thought of as expressing conceivability, for deontic logics it indicates ideal worlds, for temporal logics the relation expresses an ordering in time, etc. The models that we use to interpret formulae from L^{X} contain a set S of possible worlds, representing actual and hypothetical states of affairs, a valuation π on the elements of Π, indicating which atomic propositions are true in which possible world, a relation R denoting epistemic accessibility, and two functions $\mathtt{r}_0$ and $\mathtt{c}_0$ dealing with (the result, opportunity and ability for) atomic actions.

2.6. Definition. A model M for L^{X} is a tuple containing at least the following elements:

- a non-empty set S of possible worlds or states.
- a valuation $\pi : \Pi \times \mathrm{S} \to \mathrm{bool}$ on propositional symbols.
- a function $\mathrm{R} : \mathrm{A} \to \wp(\mathrm{S} \times \mathrm{S})$ indicating the epistemic alternatives of agents. This function is demanded to be such that $\mathrm{R}(i)$ is an equivalence relation for all $i \in \mathrm{A}$.
- a function $\mathtt{r}_0 : \mathrm{A} \times \mathrm{At} \to \mathrm{S} \to \mathrm{S}^{\cdot}$ indicating the state-transitions caused by the execution of atomic actions.
- a function $\mathtt{c}_0 : \mathrm{A} \times \mathrm{At} \to \mathrm{S} \to \mathrm{bool}$ determining the abilities of agents with regard to atomic actions.

The class containing all models for L^{X} is denoted by $\mathbf{M}^{\mathrm{X}}$. The letter M, possibly marked, denotes a typical model, and s, t, u, possibly marked, are used as typical elements of the set of states.

The relation $\mathrm{R}(i)$ indicates which pairs of worlds are indistinguishable for agent i on the basis of its knowledge: if $(s, s') \in \mathrm{R}(i)$ then whenever s is the description of the actual world, s' might as well be for all agent i knows. To ensure that knowledge indeed

has the properties sketched in the previous section, it is demanded that $\mathrm{R}(i)$ is an equivalence relation for all i, i.e. $\mathrm{R}(i)$ is reflexive ($(s,s) \in \mathrm{R}(i)$) and Euclidean (if $(s,s') \in \mathrm{R}(i)$ and $(s,s'') \in \mathrm{R}(i)$ then also $(s',s'') \in \mathrm{R}(i)$)[4]. That this demand ensures that knowledge behaves as desired is stated in Proposition 2.8 and explained in Proposition 2.16. The function $\mathtt{r}_0$ characterises occurrences of atomic events, i.e. events consisting of an agent performing an atomic action: whenever s is some possible world, then $\mathtt{r}_0(i,a)(s)$ represents the state of affairs following execution of the atomic action a in the possible world s by the agent i. Since atomic actions are inherently deterministic, $\mathtt{r}_0(i,a)(s)$ yields at most one state of affairs as the one resulting from the occurrence of the event $\mathrm{do}_i(a)$ in s. If $\mathtt{r}_0(i,a)(s) = \emptyset$, we will sometimes say that execution of a by i in s leads to the (unique) *counterfactual state of affairs*, i.e. a state of affairs which is neither actual nor hypothetical, but counterfactual. The function $\mathtt{c}_0$ acts as a kind of valuation on atomic actions, i.e. $\mathtt{c}_0(i,a)(s)$ indicates whether agent i has the ability to perform the action a in the possible world s.

Formulae from the language L^{X} are interpreted on the possible worlds in the models from $\mathbf{M}^{\mathrm{X}}$. Due to the fact that interpretations of formulae concerning opportunities, results and abilities differ with the class of actions that is considered, it is not possible to give a uniform interpretation of these formulae that is suitable for all the systems defined in this thesis. Formulae that are interpreted in exactly the same way in all the logical systems that we present are called *core formulae*. Of these, propositional symbols are directly interpreted using the valuation π: a propositional symbol p is true in a state s iff $\pi(p,s)$ yields the value $\mathbf{1}$. Negations and conjunctions are interpreted as in classical logic: a formula $\neg\varphi$ is true in a state s iff φ is not true in s and $\varphi \wedge \psi$ is true in s iff both φ and ψ are true in s. The knowledge formulae $\mathbf{K}_i\varphi$ are interpreted using the epistemic accessibility relation $\mathrm{R}(i)$: agent i knows that φ in s iff φ is true in all the possible worlds that the agent considers epistemically compatible with s.

2.7. Definition. The binary relation $\models^{\mathrm{X}}$ between a formula and a pair M, s consisting of a model M and a state s in M is for the core formulae defined by

$$
\begin{array}{ll}
\mathrm{M}, s \models^{\mathrm{X}} p & \Leftrightarrow \pi(p,s) = \mathbf{1} \text{ for } p \in \Pi \\
\mathrm{M}, s \models^{\mathrm{X}} \neg\varphi & \Leftrightarrow \text{not } (\mathrm{M}, s \models^{\mathrm{X}} \varphi) \\
\mathrm{M}, s \models^{\mathrm{X}} \varphi \wedge \psi & \Leftrightarrow \mathrm{M}, s \models^{\mathrm{X}} \varphi \text{ and } \mathrm{M}, s \models^{\mathrm{X}} \psi \\
\mathrm{M}, s \models^{\mathrm{X}} \mathbf{K}_i\varphi & \Leftrightarrow \forall s' \in \mathrm{S}((s,s') \in \mathrm{R}(i) \Rightarrow \mathrm{M}, s' \models^{\mathrm{X}} \varphi)
\end{array}
$$

The formula φ is $\models^{\mathrm{X}}$-satisfiable in the model M iff $\mathrm{M}, s \models^{\mathrm{X}} \varphi$ for some s in M; φ is $\models^{\mathrm{X}}$-valid in M, denoted by $\mathrm{M} \models^{\mathrm{X}} \varphi$, iff $\mathrm{M}, s \models^{\mathrm{X}} \varphi$ for all s in M. The formula φ is $\models^{\mathrm{X}}$-satisfiable in $\mathbf{M}^{\mathrm{X}}$ iff φ is $\models^{\mathrm{X}}$-satisfiable in some $\mathrm{M} \in \mathbf{M}^{\mathrm{X}}$; φ is $\models^{\mathrm{X}}$-valid in $\mathbf{M}^{\mathrm{X}}$,

[4] The formulation of equivalence relations as being reflexive and Euclidean is equivalent to the more standard one of reflexivity, transitivity and symmetry while providing some technical advantages.

denoted by $\models^{\mathrm{X}} \varphi$, iff φ is $\models^{\mathrm{X}}$-valid in all $\mathrm{M} \in \mathbf{M}^{\mathrm{X}}$. Whenever $\models^{\mathrm{X}}$ is clear from the context, we drop it as a prefix and simply speak of a formula φ being satisfiable or valid in a (class of) model(s). For a given model M, we define $[s]_{\mathrm{R}(i)} \triangleq \{s' \in \mathrm{S} \mid (s, s') \in \mathrm{R}(i)\}$ and $[\![\varphi]\!]_{\mathrm{M}} \triangleq \{s \in \mathrm{S} \mid \mathrm{M}, s \models \varphi\}$. Whenever the model M is clear from the context, the latter notion is usually simplified to $[\![\varphi]\!]$.

When demanding the agents' epistemic accessibility relations to be equivalence relation, the modal operator $\mathbf{K}$ indeed formalises the notion of knowledge discussed in the previous section.

2.8. Proposition. *For all $i \in \mathrm{A}$ and $\varphi, \psi \in \mathrm{L}^{\mathrm{X}}$ we have:*

1. $\models^{\mathrm{X}} \mathbf{K}_i(\varphi \to \psi) \to (\mathbf{K}_i\varphi \to \mathbf{K}_i\psi)$ *K*
2. $\models^{\mathrm{X}} \varphi \Rightarrow \models^{\mathrm{X}} \mathbf{K}_i\varphi$ *N*
3. $\models^{\mathrm{X}} \mathbf{K}_i\varphi \to \varphi$ *T*
4. $\models^{\mathrm{X}} \mathbf{K}_i\varphi \to \mathbf{K}_i\mathbf{K}_i\varphi$ *4*
5. $\models^{\mathrm{X}} \neg\mathbf{K}_i\varphi \to \mathbf{K}_i\neg\mathbf{K}_i\varphi$ *5*

The first two items of Proposition 2.8 formalise that $\mathbf{K}_i$ is a normal modal operator: $\mathbf{K}_i$ satisfies both the K-axiom and the necessitation rule N (the names of these and other modal axioms are according to the Chellas classification [19]). Furthermore, $\mathbf{K}_i$ satisfies the axioms of veridicality (the T-axiom), positive introspection (axiom 4) and negative introspection (axiom 5).

2.3 A menagerie of definitions

In this section we present a uniform account of some definitions that will be used at various places throughout the rest of this thesis.

2.3.1 Actions from a syntactical point of view

Given the inductive definition of actions, it may come as no surprise that composite actions can be unravelled into the sequences of atomic actions and confirmations — we will use the general term *semi-atomic* actions for actions that are either atomic or confirmations — that constitute them. From the syntax of an action alone it is possible to determine a set of sequences of semi-atomic actions, called *finite computation sequences*, that may possibly occur in a halting execution of the action. Since the way in which execution of an action evolves depends on the state of affairs in which the action is performed, it is obvious that in general not all finite computation sequences of an action will actually occur when the action is performed. However, it can easily be shown that the sequence of semi-atomic actions that actually constitutes the execution of a composite action is always a member of the set of finite computation sequences of the action.

2.9. Definition. The class $\mathrm{confirm}(\mathrm{L}^{\mathrm{X}})$ of confirmations based on L^{X} is defined by $\mathrm{confirm}(\mathrm{L}^{\mathrm{X}}) \triangleq \{\mathtt{confirm}\,\varphi \mid \varphi \in \mathrm{L}^{\mathrm{X}}\}$. The class $\mathrm{Ac}_{\mathrm{s}}^{\mathrm{X}}$ of semi-atomic actions is defined by $\mathrm{Ac}_{\mathrm{s}}^{\mathrm{X}} \triangleq \mathrm{At} \cup \mathrm{confirm}(\mathrm{L}^{\mathrm{X}})$. The class $\mathrm{Ac}_{\mathrm{b}}^{\mathrm{X}}$ of basic actions based on L^{X} is the smallest superset of $\mathrm{Ac}_{\mathrm{s}}^{\mathrm{X}}$ closed under sequential composition.

2.10. Definition. The function CS^{X}, yielding the set of finite computation sequences of a given action, is inductively defined as follows:

$$
\begin{array}{ll}
\mathrm{CS}^{\mathrm{X}} & : \ \mathrm{Ac}^{\mathrm{X}} \to \wp(\mathrm{Ac}_{\mathrm{b}}^{\mathrm{X}}) \\
\mathrm{CS}^{\mathrm{X}}(\alpha) & = \{\alpha\} \text{ if } \alpha \in \mathrm{Ac}_{\mathrm{s}}^{\mathrm{X}} \\
\mathrm{CS}^{\mathrm{X}}(\alpha_1;\alpha_2) & = \{\alpha_1';\alpha_2' \mid \alpha_1' \in \mathrm{CS}^{\mathrm{X}}(\alpha_1), \alpha_2' \in \mathrm{CS}^{\mathrm{X}}(\alpha_2)\} \\
\mathrm{CS}^{\mathrm{X}}(\mathtt{if}\,\varphi\,\mathtt{then}\,\alpha_1\,\mathtt{else}\,\alpha_2\,\mathtt{fi}) & = \mathrm{CS}^{\mathrm{X}}(\mathtt{confirm}\,\varphi;\alpha_1) \cup \mathrm{CS}^{\mathrm{X}}(\mathtt{confirm}\,\neg\varphi;\alpha_2) \\
\mathrm{CS}^{\mathrm{X}}(\mathtt{while}\,\varphi\,\mathtt{do}\,\alpha\,\mathtt{od}) & = \cup_{k=1}^{\infty} \mathit{Seq}_k(\mathtt{while}\,\varphi\,\mathtt{do}\,\alpha\,\mathtt{od}) \cup \{\mathtt{confirm}\,\neg\varphi\}
\end{array}
$$

where for $k \geq 1$

$$
\begin{array}{ll}
\mathit{Seq}_k(\mathtt{while}\,\varphi\,\mathtt{do}\,\alpha\,\mathtt{od}) & = \{\Pi_{j=1}^{k}(\varphi,\alpha_j') \mid \alpha_j' \in \mathrm{CS}^{\mathrm{X}}(\alpha) \text{ for } j = 1,\ldots,k\} \\
\Pi_{j=1}^{l}(\psi,\beta_j) & = (\mathtt{confirm}\,\psi;\beta_1);\ldots;(\mathtt{confirm}\,\psi;\beta_l);\mathtt{confirm}\,\neg\psi
\end{array}
$$

On several occasions it turns out to be handy to have the possibility to manipulate initial fragments of basic actions. To this end we introduce the binary relation $\mathrm{Prefix}^{\mathrm{X}}$, which applies to pairs of basic actions such that the first one is an initial fragment of the second one. Defining semi-atomic actions to be of length 1, we define the length of a basic action to be the number of semi-atomic actions that constitute it. Using the $\mathrm{Prefix}^{\mathrm{X}}$ relation and the function $|\_|^{\mathrm{X}}$ which determines the length of basic actions, we can define a function which yields the prefix of a given length of a basic action.

2.11. Definition. The relation $\mathrm{Prefix}^{\mathrm{X}} \subseteq \mathrm{Ac}_{\mathrm{b}}^{\mathrm{X}} \times \mathrm{Ac}_{\mathrm{b}}^{\mathrm{X}}$ is defined to be the smallest superset of $\{(\alpha,\alpha) \mid \alpha \in \mathrm{Ac}_{\mathrm{b}}^{\mathrm{X}}\}$ such that for all $\alpha,\beta_1,\beta_2 \in \mathrm{Ac}_{\mathrm{b}}^{\mathrm{X}}$ it holds that:

$$
\begin{array}{ll}
\mathrm{Prefix}^{\mathrm{X}}(\alpha,\beta_1) & \Rightarrow \mathrm{Prefix}^{\mathrm{X}}(\alpha,\beta_1;\beta_2) \\
\mathrm{Prefix}^{\mathrm{X}}(\alpha;\beta_1,\alpha;\beta_2) & \Leftrightarrow \mathrm{Prefix}^{\mathrm{X}}(\beta_1,\beta_2)
\end{array}
$$

2.12. Definition. The function $|\_|^{\mathrm{X}}$ and the partial function $|\_|_{\_}^{\mathrm{X}}$, yielding the length of a basic action and the initial fragment of a basic action of a particular length, respectively, are defined as follows.

$$
\begin{array}{lll}
|\_|^{\mathrm{X}} & : & \mathrm{Ac}_{\mathrm{b}}^{\mathrm{X}} \to \mathbb{N} \\
|\alpha|^{\mathrm{X}} & = & 1 \text{ if } \alpha \in \mathrm{Ac}_{\mathrm{s}}^{\mathrm{X}} \\
|\alpha_1;\alpha_2|^{\mathrm{X}} & = & |\alpha_1|^{\mathrm{X}} + |\alpha_2|^{\mathrm{X}}
\end{array}
$$

$$
\begin{array}{lll}
|\_|_{\_}^{\mathrm{X}} & : & \mathrm{Ac}_{\mathrm{b}}^{\mathrm{X}} \times \mathbb{N} \to \mathrm{Ac}_{\mathrm{b}}^{\mathrm{X}} \\
|\alpha|_n^{\mathrm{X}} \text{ is undefined} & \Leftrightarrow & n > |\alpha|^{\mathrm{X}} \\
|\alpha|_n^{\mathrm{X}} = \alpha' & \Leftrightarrow & n \leq |\alpha|^{\mathrm{X}} \,\&\, \mathrm{Prefix}^{\mathrm{X}}(\alpha',\alpha) \,\&\, |\alpha'|^{\mathrm{X}} = n
\end{array}
$$

2.3.2 Schemas, frames and correspondences

To investigate properties of knowledge and actions, it will often prove useful to refer to *schemas*, which are sets of formulae, usually of a particular form. Using schemas one may abstract from particular agents and particular formulae, thereby having the possibility to formulate certain qualities of knowledge and action in a very general way. Due to the expressiveness of our multi-modal language, our definition of schema has to be slightly more elaborate than the usual ones [19, 43], in the sense that we specify explicitly on which parameters the elements of a schema are allowed to differ. Usually when defining schemas there is only one possibility for this varying parameter.

2.13. Definition. If $\varphi \in \mathrm{L}^{\mathrm{X}}$ is some formula then the schema φ in $x_1 \in X_1, \ldots, x_m \in X_m$ consists of all formulae from L^{X} that result from uniformly replacing x_k by an arbitrary element of X_k. For example, for $3 \in \mathrm{A}$ and $q \in \Pi$ the formula $\langle \mathrm{do}_3(a) \rangle q$ is an element of the schema $\langle \mathrm{do}_i(a) \rangle f$ in $i \in \mathrm{A}, f \in \mathrm{L}_0$. If $x_1, \ldots, x_m$ are either understood or irrelevant we simply refer to φ as a schema.

Where schemas are used to express general properties of knowledge and actions on the syntactic level, *frames* can be used to do so on the semantic level. Informally speaking, a frame can be seen as a model without a valuation. By leaving out the valuation one may abstract from particular properties of knowledge and actions that are due to the valuation rather than inherently due to the nature of knowledge and/or action itself. Truth in a frame is defined in terms of truth in all models that can be constructed by adding a valuation to the frame.

2.14. Definition. A frame F for a model $\mathrm{M} \in \mathbf{M}^{\mathrm{X}}$ is a tuple consisting of the elements of M except for the valuation π. The class of all frames for models from $\mathbf{M}^{\mathrm{X}}$ is denoted by $\mathbf{F}^{\mathrm{X}}$. If $\mathrm{F} \in \mathbf{F}^{\mathrm{X}}$ is some frame then (F, π) denotes the model generated by the elements of F and the valuation π. For $\mathrm{F} \in \mathbf{F}^{\mathrm{X}}$ and $\varphi \in \mathrm{L}^{\mathrm{X}}$ we define

- $\mathrm{F} \models^{\mathrm{X}} \varphi \Leftrightarrow (\mathrm{F}, \pi) \models^{\mathrm{X}} \varphi$ for all valuations π
- $\mathbf{F}^{\mathrm{X}} \models^{\mathrm{X}} \varphi \Leftrightarrow \mathrm{F} \models^{\mathrm{X}} \varphi$ for all $\mathrm{F} \in \mathbf{F}^{\mathrm{X}}$

Since schemas are used to express general properties of knowledge and action syntactically, and frames can be used to do this semantically, the question arises as to how these notions relate. In particular, it is both interesting and important to try to single out constraints on frames that exactly correspond to certain properties of knowledge and/or action, expressed in the form of schemas. The area of research called *correspondence theory* deals with finding relations — *correspondences* — between schemas and (first-order expressible) constraints on frames. A schema is said to correspond to a constraint on frames if the schema is satisfied in a frame iff the frame obeys the constraint.

2.15. Definition. If φ is a schema and $\mathrm{F} \in \mathbf{F}^{\mathrm{X}}$ is some frame then $\mathrm{F} \models^{\mathrm{X}} \varphi$ iff $\mathrm{F} \models^{\mathrm{X}} \psi$ for all formulae ψ that are in the schema φ. If P is a formula in the first-order language containing the constants $\mathbf{0}, \mathbf{1}$, the functions $\mathrm{R}, \mathtt{r}_0, \mathtt{c}_0$ and equality such that the free variables of P are among $x_1, \ldots, x_m$ then $\mathrm{F} \models^{\mathrm{fo}} P$ iff F satisfies P for all $x_1, \ldots, x_m$. The schema φ corresponds to the first-order formula P, notation $\varphi \sim^{\mathrm{X}} P$ iff $\forall F \in \mathbf{F}^{\mathrm{X}}(F \models^{\mathrm{X}} \varphi \Leftrightarrow F \models^{\mathrm{fo}} P)$.

As already hinted at above, the properties that we demand knowledge to obey correspond to constraints on the epistemic accessibility relations $\mathrm{R}(i)$. In Definition 2.6 we demanded these relations to be equivalence relations, and this demand indeed corresponds to knowledge being veridical and satisfying the properties of positive and negative introspection. The proof of the following proposition is standard and well-known from the literature [61, 96].

2.16. Proposition. *The following correspondences hold.*

1. $\mathrm{T} \sim^{\mathrm{X}} \forall s((s,s) \in \mathrm{R}(i))$, *i.e.* $\mathrm{R}(i)$ *is reflexive*
2. $4 \sim^{\mathrm{X}} \forall s, s', s''((s,s') \in \mathrm{R}(i) \,\&\, (s',s'') \in \mathrm{R}(i) \Rightarrow (s,s'') \in \mathrm{R}(i))$, *i.e.* $\mathrm{R}(i)$ *is transitive*
3. $5 \sim^{\mathrm{X}} \forall s, s', s''((s,s') \in \mathrm{R}(i) \,\&\, (s,s'') \in \mathrm{R}(i) \Rightarrow (s',s'') \in \mathrm{R}(i))$, *i.e.* $\mathrm{R}(i)$ *is Euclidean*

On occasion it will turn out to be handy to single out subclasses of models that have certain properties. These subclasses can be defined as the models that are generated by a class of frames that satisfy certain constraints.

2.17. Definition. If $P_1, \ldots, P_m$ are (the names of) first-order expressible constraints on the elements of the frames from $\mathbf{F}^{\mathrm{X}}$, then $\mathbf{F}^{\mathrm{X}}_{P_{k_1}, \ldots, P_{k_l}}$ contains all frames from $\mathbf{F}^{\mathrm{X}}$ that satisfy the constraints $P_{k_1}, \ldots, P_{k_l}$. The class $\mathbf{M}^{\mathrm{X}}_{P_{k_1}, \ldots, P_{k_l}}$ contains all models that are generated by the elements of $\mathbf{F}^{\mathrm{X}}_{P_{k_1}, \ldots, P_{k_l}}$.

2.3.3 Additional properties of actions

The core of the formal languages considered in this thesis, as embodied by the language L^{X}, is already sufficiently expressive to formalise various properties of knowledge, actions and their interplay. The first of the properties that we consider here is *accordance*. Informally speaking, accordant actions are known to behave according to plan, i.e. for an accordant action it will be the case that things that an agent expects — on the basis of its knowledge — to hold in the future state of affairs that will result from it executing the action, are indeed known to be true by the agent when that future state of affairs has been brought about. Accordance of actions may be an important property in the context of agents planning to achieve certain goals. For if the agent knows (now) that performing some accordant action will bring about some goal, then it will be satisfied after it has

executed the action: the agent knows that the goal is brought about. From a formal point of view, accordance of an action α corresponds to the schema $\mathbf{K}_i[\mathrm{do}_i(\alpha)]\varphi \to [\mathrm{do}_i(\alpha)]\mathbf{K}_i\varphi$ in $i \in \mathrm{A}$ and $\varphi \in \mathrm{L}^{\mathrm{X}}$.

The notion of *determinism* was already touched upon at the introduction of atomic actions, where it was stated that having a unique outcome is an essential characteristic of the 'atomicness' of actions. With the exception of those introduced in Chapter 4, all actions that we consider turn out to be deterministic. The notion of determinism of an action α is formalised through the schema $\langle \mathrm{do}_i(\alpha)\rangle\varphi \to [\mathrm{do}_i(\alpha)]\varphi$ in $i \in \mathrm{A}$ and $\varphi \in \mathrm{L}^{\mathrm{X}}$.

Whenever an action is *idempotent*, consecutively executing the action twice — or in general an arbitrary number of times — will have exactly the same results as performing the action just once. In a sense, the state of affairs reached after the first performance of the action can be seen as a kind of fixed-point of execution of the action. The most simple idempotent action in our framework is the void action `skip`: performing it once, twice or an arbitrary number of times will not affect the state of affairs in any way whatsoever. Other examples of idempotent actions are the informative actions encountered in Chapter 5. Formally, idempotence of an action α corresponds to the schema $[\mathrm{do}_i(\alpha;\alpha)]\varphi \leftrightarrow [\mathrm{do}_i(\alpha)]\varphi$, or equivalently $\langle \mathrm{do}_i(\alpha;\alpha)\rangle\varphi \leftrightarrow \langle \mathrm{do}_i(\alpha)\rangle\varphi$, in $i \in \mathrm{A}$ and $\varphi \in \mathrm{L}^{\mathrm{X}}$.

Agents always have the opportunity to perform *realisable* actions, regardless of the circumstances, i.e. there never is an external factor that may prevent or interfere with the performance of such an action. Realisable actions will in general not depend on, interfere with, or affect the external environment of the performing agent, but instead have some agent-internal effects, like contributing to the agent's information. The actions modelling observations that we encounter in Chapter 5 turn out to be realisable. The property of *A-realisability* relates ability and opportunity. For actions that are A-realisable, ability implies opportunity, i.e. whenever an agent is able to perform the action it automatically has the opportunity to perform it. Realisable actions are trivially A-realisable, and so are actions that no agent is ever capable of performing, but it seems hard to think of non-trivial examples of regular, mundane actions that an agent is able to execute and therefore automatically has the opportunity to do so. One example of a special, non-regular action that is A-realisable is the one that we introduce in Chapter 5 to model the jumping to conclusions which constitutes default reasoning. Although we do not think the property of A-realisability to be desirable in general, it seems to be a reasonable option when one wants to correlate abilities and opportunities more stringently than is done in our account. In any case it is far more reasonable to assume that ability implies opportunity than the reverse, given the fact that 'abilities are states that are acquired with effort [whereas] opportunities are there for the taking until they pass' ([68], p. 133). Realisability of an action α is formalised through the schema $\langle \mathrm{do}_i(\alpha)\rangle\top$ in $i \in \mathrm{A}$; A-realisability corresponds to $\mathbf{A}_i\alpha \to \langle \mathrm{do}_i(\alpha)\rangle\top$ as a schema in $i \in \mathrm{A}$.

The following definition summarises the properties discussed above in a formal way.

2.18. Definition. Let $\alpha \in \mathrm{Ac}^{\mathrm{X}}$ be some action and let $\mathrm{F} \in \mathbf{F}^{\mathrm{X}}$. The right-hand side of the following definitions is to be understood as a schema in $i \in \mathrm{A}, \varphi \in \mathrm{L}^{\mathrm{X}}$.

- α is accordant in F iff $\mathrm{F} \models^{\mathrm{X}} \mathbf{K}_i[\mathrm{do}_i(\alpha)]\varphi \to [\mathrm{do}_i(\alpha)]\mathbf{K}_i\varphi$
- α is deterministic in F iff $\mathrm{F} \models^{\mathrm{X}} \langle \mathrm{do}_i(\alpha)\rangle\varphi \to [\mathrm{do}_i(\alpha)]\varphi$
- α is idempotent in F iff $\mathrm{F} \models^{\mathrm{X}} [\mathrm{do}_i(\alpha;\alpha)]\varphi \leftrightarrow [\mathrm{do}_i(\alpha)]\varphi$
- α is realisable in F iff $\mathrm{F} \models^{\mathrm{X}} \langle \mathrm{do}_i(\alpha)\rangle\top$
- α is A-realisable in F iff $\mathrm{F} \models^{\mathrm{X}} \mathbf{A}_i\alpha \to \langle \mathrm{do}_i(\alpha)\rangle\top$

If Prop is any of the properties defined above, we say that α has the property Prop in $\mathbf{F}^{\mathrm{X}}$ iff α has the property Prop in every $\mathrm{F} \in \mathbf{F}^{\mathrm{X}}$.

2.4 Summary and conclusions

In this chapter we presented the formal framework which constitutes the core of all formal systems presented in this thesis. After a somewhat philosophical exposition on knowledge, actions and events, we presented the core language L^{X}. This language is a propositional, multi-modal, exogenous language, containing modalities representing knowledge, opportunity and result, and an operator formalising ability. The models that are used to interpret formulae from the language L^{X} are Kripke-style possible worlds models. These models interpret knowledge by means of an accessibility relation on worlds; opportunity, result and ability are interpreted using designated functions. Having introduced the basic formal framework, we presented a variety of definitions, among which those of schemas, frames and correspondences, and functions to unravel actions. We furthermore considered various properties of actions, which we formalised using schemas and frames.

2.4.1 Bibliographical notes

Since this chapter provides but the core of the formal frameworks presented in this thesis, this is hardly the place for extensive bibliographical notes. However, we would like to spend some words on our exogenous account of ability and its place in the literature. In the literature on philosophical logic, several formalisations of ability have been proposed, but few of these are exogenous. Approaches like those of Brown [10] and Elgesem [29] use modal operators, ranging over formulae, to formalise ability: $\mathbf{A}_i\varphi$ is then to be read as 'agent i is able to bring about circumstances in which φ is true'. This reading of ability seems a bit artificial, and that in itself would be enough reason to pursue an exogenous approach. Moreover, endogenous approaches may suffer from the problems of logical omniscience [31, 96] that are known to plague normal modal operators (in a slightly different form we will encounter these problems in Chapter 6). The only formal system that we know of in which it is hinted at an exogenous approach towards ability is the one proposed by Penther [102]. Penther (independently) proposed a dynamic logic of action,

which is a slightly refined version of the action fragment of our framework, and suggested a formalisation of abilities. Conceptually, Penther's account is inspired by Kenny's ideas on abilities and opportunities, and therefore seems very similar to the one that we present. Penther claims that her main interest (just like ours) is in the *compositional* behaviour of ability rather than its actual nature. It is therefore remarkable that in her formal framework she restricts herself to abilities for *atomic* actions, and does not bother to extend this to composite actions. For these atomic actions, Penther presupposes the property of A-realisability (obviously without using this term), which in our opinion binds the notions of ability and opportunity too tight.

Chapter 3

A logic of capabilities

> *It seems that, when one is learning a new skill, be it walking or driving a car, initially one must think through each action in detail, and the cerebrum is in control; but when the skill has been mastered — and has become 'second nature' — it is the cerebellum that takes over. Moreover, it is a familiar experience that if one* thinks about *one's actions in a skill that has been so mastered, then one's easy control may be temporarily lost.* Thinking *about it seems to involve the reintroduction of cerebral control and, although a consequent flexibility of activity is thereby introduced, the flowing and precise cerebellar action is lost. No doubt such descriptions are oversimplified, but they give a reasonable flavour of the cerebellum's role.*
>
> Roger Penrose, *'The Emperor's New Mind'*.

In this chapter we present the first two of the formal systems built on the core defined in the previous chapter. The language common to both systems defined here is built up from the core clauses only, and the models used to interpret this language contain nothing but the core elements. The two systems differ in their interpretation of dynamic formulae $\langle \mathrm{do}_i(\alpha) \rangle \varphi$ and ability formulae $\mathbf{A}_i \alpha$, and in particular on their treatment of abilities in the counterfactual state of affairs. We show how some of the additional properties of action introduced in the previous chapter can be brought about by imposing suitable constraints on the frames under consideration. Moreover, we consider the notion of practical possibility, and formalise part of the reasoning of agents on the correctness and feasibility of their actions. Two slightly different proof systems are presented that are sound and complete with respect to the notions of validity associated with the two interpretations. The most striking feature of these proof systems is the use of infinitary

proof rules, i.e. rules with an infinite number of premises. We conclude this chapter with the usual summary, discuss some alternative definitions of ability for sequential compositions, provide some bibliographical notes that put our work in the context of previous research on formalisations of aspects of agency, and prove some selected propositions. Among the proofs that are given is a fairly elaborate and complicated one showing the soundness and completeness of the proof systems with respect to the appropriate notions of validity.

3.1 The KARO-architecture

The two systems that we consider here are the most basic ones of those investigated in this thesis. They share the same language and the same class of models, but differ in the interpretation of dynamic and ability formulae, and as a consequence of this, also in their proof systems. To emphasise the similarity between the two systems, we simply define one language L and one class $\mathbf{M}$ of models, without bothering to decorate them with superscripts, while on the other hand we define two different satisfiability relations and proof systems. The systems thus defined are occasionally referred to as belonging to the family of KARO-architectures, for Knowledge-Ability-Result-Opportunity-architecture[1].

The language L and its associated class Ac of actions, which are common to the two systems, are built up by the core clauses only.

3.1. Definition. The language L and the class Ac of actions are the smallest sets closed under the core clauses.

The models that are used to interpret the formulae from L (in two different ways) contain nothing but the core elements.

3.2. Definition. A model M for the language L is a tuple consisting of the core elements $\mathrm{S}, \pi, \mathrm{R}, \mathtt{r}_0$ and $\mathtt{c}_0$. The class of all these models is denoted by $\mathbf{M}$.

To interpret the dynamic formulae $\langle \mathrm{do}_i(\alpha) \rangle \varphi$ and the ability formulae $\mathbf{A}_i \alpha$, the functions $\mathtt{r}_0$ and $\mathtt{c}_0$ are lifted from the level of atomic actions to the level of composite actions. Informally, a formula $\langle \mathrm{do}_i(\alpha) \rangle \varphi$ is true in some possible world s, if the extension of $\mathtt{r}_0$ applied to i, α and s yields some successor state s' in which the formulae φ holds. A formula $\mathbf{A}_i \alpha$ is true in a state s if the extension of $\mathtt{c}_0$ yields the value $\mathbf{1}$ when applied to i, α and s. Before defining the extended versions of $\mathtt{r}_0$ and $\mathtt{c}_0$, we first motivate the choices underlying these extensions.

[1] The term KARO-architecture is inspired by the BDI-architecture, for Belief-Desire-Intention, of Rao & Georgeff [109].

3.1.1 Results and opportunities for composite actions

The extension of the function $\mathtt{r}_0$ as we present it is originally due to Halpern & Reif [44]. Although Halpern & Reif's logic is meant to reason about computer programs and not about agents performing actions, we argue that their definition is also adequate for our purposes. Using this definition, actions $\mathtt{confirm}\,\varphi$ are interpreted as genuine confirmations: whenever the formula φ is true in a state s, s is its own $\mathrm{do}_i(\mathtt{confirm}\,\varphi)$-successor. If φ does not hold in a possible world s, then the $\mathtt{confirm}\,\varphi$ action fails, and no successor state results. In practice this implies that (all) agents have the opportunity to confirm the truth of a certain formula iff the formula holds. Execution of such an action does not have any effects in the case that the formula that is confirmed holds, and leads to the counterfactual state of affairs if the formula does not hold[2].

Since the action $\alpha_1;\alpha_2$ is intuitively interpreted as 'α_1 followed by α_2', the transition caused by execution of an action $\alpha_1;\alpha_2$ equals the 'sum' of the transition caused by α_1 and the one caused by α_2 in the state brought about by execution of α_1. In the case that execution of α_1 leads to an empty set of states, execution of the action $\alpha_1;\alpha_2$ also leads to an empty set: there is no escape from the counterfactual state of affairs. In practice this implies that an agent has the opportunity to perform a sequential composition $\alpha_1;\alpha_2$ iff it has the opportunity to do α_1 (now), and doing α_1 results in the agent having the opportunity to do α_2. The results of performing $\alpha_1;\alpha_2$ equal the results of doing α_2, having done α_1.

Given its intuitive meaning, it is obvious that the transition caused by a conditional composition $\mathtt{if}\,\varphi\,\mathtt{then}\,\alpha_1\,\mathtt{else}\,\alpha_2\,\mathtt{fi}$ equals the one associated with α_1 in the case that φ holds and the one caused by execution of α_2 in the case that $\neg\varphi$ holds. This implies that an agent has the opportunity to perform an action $\mathtt{if}\,\varphi\,\mathtt{then}\,\alpha_1\,\mathtt{else}\,\alpha_2\,\mathtt{fi}$ if (it has the opportunity to confirm that) φ holds and it has the opportunity to do α_1, or (it has the opportunity to confirm that) $\neg\varphi$ holds and the agent has the opportunity to do α_2. The result of performing $\mathtt{if}\,\varphi\,\mathtt{then}\,\alpha_1\,\mathtt{else}\,\alpha_2\,\mathtt{fi}$ equals the result of α_1 in the case that φ holds and that of α_2 otherwise.

The definition of the extension of $\mathtt{r}_0$ for the repetitive composition is based on the

[2] Originally in dynamic logic [46, 70] these actions were referred to as tests instead of confirmations. As long as one deals with the behaviour of computer programs, the term 'test' is quite acceptable. However as soon as formalisations of (human) agents are concerned, one should be careful with using this term. The commonsense notion of test is that of an action, execution of which provides some kind of information. For example dope-tests and eye-tests are performed in order to acquire information on whether some athlete has been taking drugs, or whether someone's eyesight is adequate. The nature of this kind of tests is not captured by the action which just checks for the truth of some proposition, without yielding any information whatsoever. To avoid confusion we have chosen to refer to these latter kinds of actions as confirmations. Based on similar arguments, Segerberg suggested in one of his lectures to use the term 'verification' rather than 'test' when referring to actions that, like our confirmations, just check for the truth of some proposition.

idea that execution of the action $\texttt{while}\,\varphi\,\texttt{do}\,\alpha\,\texttt{od}$ comes down to sequentially testing for the truth of φ and executing α until a state is reached in which $\neg\varphi$ holds. This behaviour of the repetitive composition can be simulated by equating the transition caused by execution of $\texttt{while}\,\varphi\,\texttt{do}\,\alpha\,\texttt{od}$ with the set of transitions caused by execution of the computation sequences of $\texttt{while}\,\varphi\,\texttt{do}\,\alpha\,\texttt{od}$, an equation well-known from dynamic logic. For deterministic while-loops $\texttt{while}\,\varphi\,\texttt{do}\,\alpha\,\texttt{od}$ it holds that at most one of the actions $\beta_k = ((\texttt{confirm}\,\varphi;\alpha)^k;\texttt{confirm}\,\neg\varphi)$ with $k \in \mathbb{N}$, has an execution which does not lead to the counterfactual state of affairs. Now if such an action β_k exists, the resulting state of execution of the while-loop is defined to be the state resulting from execution of β_k, and otherwise execution of the loop is taken to lead to the counterfactual state of affairs. Using this definition implies that an agent has the opportunity to perform a repetitive composition $\texttt{while}\,\varphi\,\texttt{do}\,\alpha\,\texttt{od}$ iff it has the opportunity to perform some finite-length sequence of confirmations for φ and $\alpha's$, with $\alpha' \in \mathrm{CS}(\alpha)$, which results in $\neg\varphi$ being true; the result of executing $\texttt{while}\,\varphi\,\texttt{do}\,\alpha\,\texttt{od}$ equals that of executing this sequence.

3.1.2 Abilities for composite actions

Whereas the extension of r_0 for composite actions is more or less standard, the extension of c_0 as determining the abilities of agents for composite actions, is not. Since we are (among) the first to give a formal, exogenous account of ability, extending the function c_0 to the class of all actions involves a couple of personal choices.

We start with motivating our definitions of ability for confirmations and conditional compositions since neither of these is really controversial: the definition of ability for confirmations is indisputable since it represents a highly personal choice (and there is no accounting for tastes), and that of the ability for the conditional composition is too obvious and natural to be questioned.

We have decided to let an agent have the ability to confirm any formula that is actually true. Since confirmations do not correspond to any actions usually performed by humans, this definition seems to be perfectly acceptable, or at least it is hard to come up with any convincing counterarguments to it. Note that this definition implies that in a situation where some proposition is true, (all) agents have both the opportunity and the ability to confirm this proposition. In particular it is the case that confirmations are A-realisable: an agent has the ability to confirm some proposition φ if and only if it has the opportunity to do so.

Let us continue with defining abilities for conditionally composed actions. For these actions, ability is defined analogously to opportunity: an agent is able to perform the action $\texttt{if}\,\varphi\,\texttt{then}\,\alpha_1\,\texttt{else}\,\alpha_2\,\texttt{fi}$ iff either it is able to confirm the condition φ and perform α_1 afterwards, or it is able to confirm the negation of the condition and perform α_2. In practice this implies that having the ability to perform an action $\texttt{if}\,\varphi\,\texttt{then}\,\alpha_1\,\texttt{else}\,\alpha_2\,\texttt{fi}$

boils down to being able to do α_1 whenever φ holds and being able to do α_2 whenever φ does not hold. In our opinion this is *the* natural way to define the ability for conditionally composed actions; it's hard to think of alternatives.

Whereas the definitions of the ability for confirmations and conditional compositions are easily explained and motivated, this is not the case for those describing the ability for sequential and repetitive compositions, even though the basic ideas underlying these definitions are perfectly clear.

Informally, having the ability to perform a sequentially composed action $\alpha_1;\alpha_2$ is defined as having the ability to do α_1 now, while being able to do α_2 as a result of having done α_1. In the case that the opportunity to perform α_1 exists, i.e. performing α_1 does not result in the counterfactual state of affairs, there is no question concerning the intuitive correctness of this definition, but things are different when this opportunity is absent. It is not clear how the abilities of agents are to be determined in the counterfactual state of affairs. Probably the most acceptable approach would be to declare the question on whether the agent is able to perform an action in the counterfactual state of affairs to be meaningless, which could be formalised by extending the set of truth-values to contain an element representing undefinedness of a proposition. Since this would necessitate a considerable complication of our classical, two-valued approach, we have chosen not to explore this avenue, which leaves us with the task of assigning a classical truth-value to the agents' abilities in the counterfactual state of affairs. In general we see two ways of doing this, the first of which would be to treat all actions equally and come up with a uniform truth value for the abilities of all agents to perform any action in the counterfactual state of affairs. This approach is relatively simply to formalise, and is in fact the one that we will pursue. The second approach would be to treat each action individually, and determine the agents' abilities through other means, like for instance assuming an agent to be in the possession of certain default, or typical, abilities. This approach is further discussed at the end of this chapter. Coming back to the first approach, it is obvious that — given that there are exactly two truth-values — two ways exist to treat all actions equally with respect to the agents' abilities in the counterfactual state of affairs. The first of these could be called an *optimistic*, or bold, approach, and states that agents are omnipotent in the counterfactual state of affairs. According to this approach, in situations where an agent does have the ability but not the opportunity to perform an action α_1 it is concluded that the agent has the ability to perform the sequential composition $\alpha_1;\alpha_2$ for arbitrary actions α_2. The second approach is a *pessimistic*, or careful one. In this approach agents are assumed to be nilpotent in counterfactual situations. Thus, in situations were an agent does have the ability but not the opportunity to perform an action α_1 it is concluded that the agent is unable to perform the sequential composition $\alpha_1;\alpha_2$ for all α_2. Note that in the case that the agent has the opportunity to do α_1, optimistic and pessimistic approaches towards the

agent's ability to do $\alpha_1;\alpha_2$ coincide. Although there is a case for both definitions, neither is completely acceptable. Consider the example of a lion in a cage, which is perfectly well capable of eating a zebra, but ideally never has the opportunity to do so. Using the first definition we would have to conclude that the lion is capable of performing the sequential composition 'eat zebra; fly to the moon', which hardly seems intuitive. Using the second definition it follows that the lion is unable to perform the action 'eat zebra; do nothing', which seems equally counterintuitive. Fortunately, the problems associated with these definitions are not really serious. For they occur only in situations where an agent has the ability but not the opportunity to perform some action. And since it is exactly the combination of opportunity and ability that is important, no unwarranted conclusions can be drawn in these situations. Henceforth, we pursue both the optimistic and the pessimistic approach; in Section 3.5 we suggest alternative approaches in which the aforementioned counterintuitive situations do not occur.

Defining abilities for while-loops is even more hazardous than for sequential compositions. Intuitively it seems a good point of departure to let an agent be able to perform a while-loop only if it is at any point during execution capable of performing the next step. However, using this intuitive definition one has to be careful not to jump to undesired conclusions in the case of an action for which execution does not terminate. It seems highly counterintuitive to declare an agent, be it artificial or not, to have the reliable ability to perform an action that goes on indefinitely. For no agent is eternal: human agents die, artificial agents break down, and after all even the lifespan of the earth and the universe is bounded. Hence agents should not be able to perform actions that take infinite time. Therefore it seems reasonable to equate the ability to perform a while-loop with the ability to perform some finite-length sequence of confirmations and actions constituting the body of the while-loop, which ends in a confirmation for the negation of the condition of the loop, analogously to the equation used in extending the function $\mathtt{r}_0$ to while-loops. Accepting this equation, it is obvious that the discussion concerning the ability of agents for sequentially composed actions also becomes relevant for the repetitive composition, i.e. also with respect to abilities for while-loops a distinction between optimistic and pessimistic agents can be made. In the case that the while-loop terminates, optimistic and pessimistic approaches coincide, but in the case that execution of the action leads to the counterfactual state of affairs, they differ. Consider the situation of an agent that up to a certain point during the execution of an action $\mathtt{while}\,\varphi\,\mathtt{do}\,\alpha\,\mathtt{od}$ has been able to perform the confirmation for φ followed by α, and now finds itself in a state where φ holds, it is able to do α but does not have the opportunity for α. An optimistic agent concludes that it would have been able to finish the finite-length sequence constituting the while-loop after the (counterfactual) execution of α, and therefore considers itself to be capable of performing the while-loop. A pessimistic agent considers itself unable to finish the sequence, and thus is unable to

perform the while-loop. The demand for finiteness of execution of the while-loop and the pessimistic view on abilities provide for a very interesting combination. For in order for an agent to be able to perform an action $\texttt{while}\,\varphi\,\texttt{do}\,\alpha\,\texttt{od}$ it has to have the opportunity to perform all the steps in the execution of $\texttt{while}\,\varphi\,\texttt{do}\,\alpha\,\texttt{od}$, possibly except for the last one. Furthermore, as as result of performing the last but one step in the execution the agent should obtain the ability to perform the last one, which is a confirmation for $\neg\varphi$. Since ability and opportunity coincide for confirmations this implies that the agent has the opportunity to confirm $\neg\varphi$, i.e. the agent has the opportunity to perform the last step in the execution of $\texttt{while}\,\varphi\,\texttt{do}\,\alpha\,\texttt{od}$. But then the agent has the opportunity to perform all the steps in the execution of the while-loop, and thus has the opportunity to perform the while-loop. Hence in the pessimistic approach the ability to perform a while-loop implies the opportunity, i.e. while-loops are A-realisable!

3.1.3 Interpreting results, opportunities and abilities

To interpret dynamic and ability formulae from L in a model M for L, the functions $\mathtt{r}_0$ and $\mathtt{c}_0$ from M are extended to deal with composite, i.e. non-atomic actions. To account for the difference between the optimistic and the pessimistic outlook on the agents' abilities, we define two different extensions of $\mathtt{c}_0$, and thereby also two different interpretations. The optimistic and the pessimistic approach coincide in their extension of $\mathtt{r}_0$, but differ in the extension of $\mathtt{c}_0$ for sequentially composed actions, and hence also in their treatment of ability for repetitive compositions. The following definition presents the extensions of $\mathtt{r}_0$ and $\mathtt{c}_0$. Here functions with the superscript $\mathbf{1}$ correspond to the optimistic view, and those with the superscript $\mathbf{0}$ to the pessimistic view on the agents' abilities in the counterfactual state of affairs.

3.3. Definition. For $\mathbf{b} \in \mathrm{bool}$ we inductively define the binary relation $\models^{\mathbf{b}}$ between a formula from L and a pair M, s consisting of a model M for L and a state s in M for the dynamic and ability formulae as follows:

$$\begin{array}{ll} \mathrm{M}, s \models^{\mathbf{b}} \langle \mathrm{do}_i(\alpha)\rangle\varphi \Leftrightarrow \exists s' \in \mathrm{S}(s' = \mathtt{r}^{\mathbf{b}}(i,\alpha)(s)\,\&\,\mathrm{M}, s' \models^{\mathbf{b}} \varphi) \\ \mathrm{M}, s \models^{\mathbf{b}} \mathbf{A}_i\alpha \quad \Leftrightarrow \mathtt{c}^{\mathbf{b}}(i,\alpha)(s) = \mathbf{1} \end{array}$$

where $\mathtt{r}^{\mathbf{b}}$ and $\mathtt{c}^{\mathbf{b}}$ are defined by:

$$\begin{array}{lcl} \mathtt{r}^{\mathbf{b}} & : & \mathrm{A} \times \mathrm{Ac} \to \mathrm{S}^{\cdot} \to \mathrm{S}^{\cdot} \\ \mathtt{r}^{\mathbf{b}}(i,a)(s) & = & \mathtt{r}_0(i,a)(s) \\ \mathtt{r}^{\mathbf{b}}(i,\texttt{confirm}\,\varphi)(s) & = & s \text{ if } \mathrm{M}, s \models^{\mathbf{b}} \varphi \\ & = & \emptyset \text{ otherwise} \\ \mathtt{r}^{\mathbf{b}}(i,\alpha_1;\alpha_2)(s) & = & \mathtt{r}^{\mathbf{b}}(i,\alpha_2)(\mathtt{r}^{\mathbf{b}}(i,\alpha_1)(s)) \\ \mathtt{r}^{\mathbf{b}}(i,\texttt{if}\,\varphi\,\texttt{then}\,\alpha_1\,\texttt{else}\,\alpha_2\,\texttt{fi})(s) & = & \mathtt{r}^{\mathbf{b}}(i,\alpha_1)(s) \text{ if } \mathrm{M}, s \models^{\mathbf{b}} \varphi \end{array}$$

$$
\begin{array}{lll}
 & = & \mathtt{r}^{\mathbf{b}}(i,\alpha_2)(s) \text{ otherwise} \\
\mathtt{r}^{\mathbf{b}}(i,\mathtt{while}\,\varphi\,\mathtt{do}\,\alpha\,\mathtt{od})(s) & = & s' \text{ if } s' = \mathtt{r}^{\mathbf{b}}(i,(\mathtt{confirm}\,\varphi;\alpha)^k;\mathtt{confirm}\,\neg\varphi)(s) \\
 & & \quad \text{for some } k \in \mathbb{N} \\
 & = & \emptyset \text{ otherwise} \\
\mathtt{r}^{\mathbf{b}}(i,\alpha)(\emptyset) & = & \emptyset \\
 & & \\
\mathtt{c}^{\mathbf{b}} & : & \mathrm{A} \times \mathrm{Ac} \to \mathrm{S}^{\cdot} \to \mathrm{bool} \\
\mathtt{c}^{\mathbf{b}}(i,a)(s) & = & \mathtt{c}_0(i,a)(s) \\
\mathtt{c}^{\mathbf{b}}(i,\mathtt{confirm}\,\varphi)(s) & = & \mathbf{1} \text{ iff } \mathrm{M},s \models^{\mathbf{b}} \varphi \\
\mathtt{c}^{\mathbf{b}}(i,\alpha_1;\alpha_2)(s) & = & \mathbf{1} \text{ iff } \mathtt{c}^{\mathbf{b}}(i,\alpha_1)(s) = \mathbf{1} \,\&\, \mathtt{c}^{\mathbf{b}}(i,\alpha_2)(\mathtt{r}^{\mathbf{b}}(i,\alpha_1)(s)) = \mathbf{1} \\
\mathtt{c}^{\mathbf{b}}(i,\mathtt{if}\,\varphi\,\mathtt{then}\,\alpha_1\,\mathtt{else}\,\alpha_2\,\mathtt{fi})(s) & = & \mathbf{1} \text{ iff } \mathtt{c}^{\mathbf{b}}(i,\mathtt{confirm}\,\varphi;\alpha_1)(s) = \mathbf{1} \text{ or} \\
 & & \quad \mathtt{c}^{\mathbf{b}}(i,\mathtt{confirm}\,\neg\varphi;\alpha_2)(s) = \mathbf{1} \\
\mathtt{c}^{\mathbf{b}}(i,\mathtt{while}\,\varphi\,\mathtt{do}\,\alpha\,\mathtt{od})(s) & = & \mathbf{1} \text{ iff } \mathtt{c}^{\mathbf{b}}(i,(\mathtt{confirm}\,\varphi;\alpha)^k;\mathtt{confirm}\,\neg\varphi)(s) = \mathbf{1} \\
 & & \quad \text{for some } k \in \mathbb{N} \\
\mathtt{c}^{\mathbf{b}}(i,\alpha)(\emptyset) & = & \mathbf{b}
\end{array}
$$

By considering the last clause of $\mathtt{c}^{\mathbf{b}}$, it is obvious that for a given model M, $\mathtt{c}^{\mathbf{1}}$ and $\mathtt{c}^{\mathbf{0}}$ are different functions. Due to the mutual dependence of $\mathtt{r}^{\mathbf{b}}$, $\mathtt{c}^{\mathbf{b}}$ and $\models^{\mathbf{b}}$, this difference also affects the other notions.

3.4. Example. Consider the language $\mathrm{L}(\Pi,\mathrm{A},\mathrm{At})$ where Π is arbitrary, i is an element of A and At contains the actions a_1, a_2. Let M be a model for $\mathrm{L}(\Pi,\mathrm{A},\mathrm{At})$ such that $\mathtt{r}_0(i,a_1)(s) = \emptyset$ and $\mathtt{c}(i,a_1)(s) = \mathbf{1}$ for some state s in M. Then it holds that:

- $\mathtt{c}^{\mathbf{1}}(i,a_1;a_2)(s) = \mathbf{1}$ and $\mathtt{c}^{\mathbf{0}}(i,a_1;a_2)(s) = \mathbf{0}$
- $\mathtt{r}^{\mathbf{1}}(i,\mathtt{confirm}\,(\mathbf{A}_i a_1;a_2))(s) = s$ and $\mathtt{r}^{\mathbf{0}}(i,\mathtt{confirm}\,(\mathbf{A}_i a_1;a_2))(s) = \emptyset$
- $\mathrm{M},s \models^{\mathbf{1}} \mathbf{A}_i a_1;a_2$ and $\mathrm{M},s \not\models^{\mathbf{0}} \mathbf{A}_i a_1;a_2$

Although $\models^{\mathbf{1}}$ differs from $\models^{\mathbf{0}}$, the compositional behaviour of actions with respect to opportunities and results is identical in the two interpretations.

3.5. Proposition. *For* $\mathbf{b} \in \mathrm{bool}$, $i \in \mathrm{A}$, $\alpha,\alpha_1,\alpha_2 \in \mathrm{Ac}$ *and* $\varphi,\psi \in \mathrm{L}$ *we have:*

1. $\models^{\mathbf{b}} \langle \mathrm{do}_i(\mathtt{confirm}\,\varphi)\rangle\psi \leftrightarrow (\varphi \wedge \psi)$
2. $\models^{\mathbf{b}} \langle \mathrm{do}_i(\alpha_1;\alpha_2)\rangle\psi \leftrightarrow \langle \mathrm{do}_i(\alpha_1)\rangle\langle \mathrm{do}_i(\alpha_2)\rangle\psi$
3. $\models^{\mathbf{b}} \langle \mathrm{do}_i(\mathtt{if}\,\varphi\,\mathtt{then}\,\alpha_1\,\mathtt{else}\,\alpha_2\,\mathtt{fi})\rangle\psi \leftrightarrow ((\varphi \wedge \langle \mathrm{do}_i(\alpha_1)\rangle\psi) \vee (\neg\varphi \wedge \langle \mathrm{do}_i(\alpha_2)\rangle\psi))$
4. $\models^{\mathbf{b}} \langle \mathrm{do}_i(\mathtt{while}\,\varphi\,\mathtt{do}\,\alpha\,\mathtt{od})\rangle\psi \leftrightarrow ((\neg\varphi \wedge \psi) \vee (\varphi \wedge \langle \mathrm{do}_i(\alpha)\rangle\langle \mathrm{do}_i(\mathtt{while}\,\varphi\,\mathtt{do}\,\alpha\,\mathtt{od})\rangle\psi))$
5. $\models^{\mathbf{b}} [\mathrm{do}_i(\alpha)](\varphi \to \psi) \to ([\mathrm{do}_i(\alpha)]\varphi \to [\mathrm{do}_i(\alpha)]\psi)$
6. $\models^{\mathbf{b}} \psi \Rightarrow \models^{\mathbf{b}} [\mathrm{do}_i(\alpha)]\psi$

Proposition 3.5 is in fact nothing but a formalisation of the intuitive ideas on results and opportunities for composite actions as expressed above. The first item states that

agents have the opportunity to confirm exactly the formulae that are true, and that no state-transition takes place as the result of such a confirmation. The second item deals with the separation of the sequential composition into its elements: an agent has the opportunity to do $\alpha_1;\alpha_2$ iff it has the opportunity to do α_1 (now) and it has the opportunity to do α_2 after α_1 has been performed. The third item states that a conditionally composed action equals its 'then'-part in the case that the condition holds, and its 'else'-part if the condition does not hold. The fourth item formalises a sort of fixed-point equation for execution of while-loops: if an agent has the opportunity to perform a while-loop then it keeps this opportunity under execution of the body of the loop as long as the condition holds. The result of performing a while-loop is also fixed under executions of the body of the loop in states where φ holds, and is determined by the propositions that are true in the first state where $\neg\varphi$ holds. Note that a validity like this one does not suffice to axiomatise the repetitive composition: although it captures the idea of while-loops representing fixed-points, it fails to force termination, i.e. this formula on its own does not guarantee that agents do not have the opportunity to bring an infinitely non-terminating while-loop to its end. In the proof systems that we present in Section 3.4 this problem is solved by including suitable proof rules guiding the repetitive composition. The last two items state the normality of $[\mathrm{do}_i(\alpha)]$.

As soon as the abilities of agents come into play, the differences between $\models^{\mathbf{1}}$ and $\models^{\mathbf{0}}$ become visible, in particular for sequential and repetitive compositions.

3.6. Proposition. *For* $\mathbf{b} \in$ bool, $i \in \mathrm{A}$, $\alpha, \alpha_1, \alpha_2 \in \mathrm{Ac}$ *and* $\varphi \in \mathrm{L}$ *we have:*

1. $\models^{\mathbf{b}} \mathbf{A}_i\mathtt{confirm}\,\varphi \leftrightarrow \varphi$
2. $\models^{\mathbf{1}} \mathbf{A}_i\alpha_1;\alpha_2 \leftrightarrow \mathbf{A}_i\alpha_1 \wedge [\mathrm{do}_i(\alpha_1)]\mathbf{A}_i\alpha_2$
3. $\models^{\mathbf{0}} \mathbf{A}_i\alpha_1;\alpha_2 \leftrightarrow \mathbf{A}_i\alpha_1 \wedge \langle\mathrm{do}_i(\alpha_1)\rangle\mathbf{A}_i\alpha_2$
4. $\models^{\mathbf{b}} \mathbf{A}_i\mathtt{if}\,\varphi\,\mathtt{then}\,\alpha_1\,\mathtt{else}\,\alpha_2\,\mathtt{fi} \leftrightarrow ((\varphi \wedge \mathbf{A}_i\alpha_1) \vee (\neg\varphi \wedge \mathbf{A}_i\alpha_2))$
5. $\models^{\mathbf{1}} \mathbf{A}_i\mathtt{while}\,\varphi\,\mathtt{do}\,\alpha\,\mathtt{od} \leftrightarrow (\neg\varphi \vee (\varphi \wedge \mathbf{A}_i\alpha \wedge [\mathrm{do}_i(\alpha)]\mathbf{A}_i\mathtt{while}\,\varphi\,\mathtt{do}\,\alpha\,\mathtt{od}))$
6. $\models^{\mathbf{0}} \mathbf{A}_i\mathtt{while}\,\varphi\,\mathtt{do}\,\alpha\,\mathtt{od} \leftrightarrow (\neg\varphi \vee (\varphi \wedge \mathbf{A}_i\alpha \wedge \langle\mathrm{do}_i(\alpha)\rangle\mathbf{A}_i\mathtt{while}\,\varphi\,\mathtt{do}\,\alpha\,\mathtt{od}))$

The first and the fourth item of Proposition 3.6 deal with the actions for which abilities are defined in a straightforward manner: agents are able to confirm exactly the true formulae, and having the ability to perform a conditional composition comes down to having the 'right' ability, dependent on the truth or falsity of the condition. The differences between the optimistic and the pessimistic outlook on abilities in the counterfactual state of affairs are clearly visible in the other items of Proposition 3.6. Optimistic agents are assumed to be omnipotent in counterfactual situations, and therefore it suffices for the agent to be able to do α_2 as a conditional result of doing α_1. A pessimistic agent needs certainty, and therefore demands to have the opportunity to do α_1 before concluding anything on its abilities following execution of α_1. This behaviour of optimistic and

pessimistic agents is formalised in the second and the third item, respectively. The fifth and sixth item formalise an analogous behaviour for repetitive compositions: optimistic agents are satisfied with conditional results (item 5) whereas pessimistic agents demand certainty (item 6). A consequence of this demand for certainty is the A-realisability of while-loops as long as the pessimistic view is taken. This property is formalised in the last item of the following proposition. The other items of Proposition 3.7 deal with the properties of confirmations. It turns out that confirmations are idempotent and A-realisable. Although these properties are not universally satisfied by all actions, it may not be too surprising that the `confirm` action, being a kind of an outsider, does have these properties.

3.7. Proposition. *For* $\mathbf{b} \in \mathrm{bool}$ *we find the following to be true. Here the formulae at the right-hand side have to be understood as schemas in* $i \in \mathrm{A}$, $\varphi, \psi \in \mathrm{L}$ *and* $\alpha \in \mathrm{Ac}$.

1. $\mathbf{F} \models^{\mathbf{b}} [\mathrm{do}_i(\mathtt{confirm}\,\varphi; \mathtt{confirm}\,\varphi)]\psi \leftrightarrow [\mathrm{do}_i(\mathtt{confirm}\,\varphi)]\psi$
2. $\mathbf{F} \models^{\mathbf{b}} \mathbf{A}_i\mathtt{confirm}\,\varphi \rightarrow \langle \mathrm{do}_i(\mathtt{confirm}\,\varphi)\rangle\top$
3. $\mathbf{F} \models^{\mathbf{0}} \mathbf{A}_i\mathtt{while}\,\varphi\,\mathtt{do}\,\alpha\,\mathtt{od} \rightarrow \langle \mathrm{do}_i(\mathtt{while}\,\varphi\,\mathtt{do}\,\alpha\,\mathtt{od})\rangle\top$

The compositional behaviour of sequential and repetitive compositions differs for the two interpretations only in situations where an agent lacks opportunities. If all appropriate opportunities are present, there is no difference for the two interpretations, a property which is formalised in the following corollary.

3.8. Corollary. *For* $i \in \mathrm{A}$, $\alpha, \alpha_1, \alpha_2 \in \mathrm{Ac}$ *and* $\varphi \in \mathrm{L}$ *we have:*

- $\models^{\mathbf{1}} \langle \mathrm{do}_i(\alpha_1)\rangle\top \rightarrow (\mathbf{A}_i\alpha_1; \alpha_2 \leftrightarrow \mathbf{A}_i\alpha_1 \wedge \langle \mathrm{do}_i(\alpha_1)\rangle\mathbf{A}_i\alpha_2)$
- $\models^{\mathbf{1}} \langle \mathrm{do}_i(\mathtt{while}\,\varphi\,\mathtt{do}\,\alpha\,\mathtt{od})\rangle\top \rightarrow$
  $(\mathbf{A}_i\mathtt{while}\,\varphi\,\mathtt{do}\,\alpha\,\mathtt{od} \leftrightarrow (\neg\varphi \vee (\varphi \wedge \mathbf{A}_i\alpha \wedge \langle \mathrm{do}_i(\alpha)\rangle\mathbf{A}_i\mathtt{while}\,\varphi\,\mathtt{do}\,\alpha\,\mathtt{od})))$

3.2 Additional properties of actions in the KARO-architecture

At the end of the previous chapter we defined various properties of actions in terms of our framework. Here we show how these properties can be brought about to hold for all actions by imposing constraints on the functions R, $\mathtt{r}_0$ and $\mathtt{c}_0$.

On the level of atomic actions, the properties introduced at the end of the previous chapter correspond to first-order expressible constraints on R, $\mathtt{r}_0$ and $\mathtt{c}_0$. In Proposition 3.10 we present the correspondences for the properties of accordance, determinism, idempotence, realisability and A-realisability, respectively. Since we have defined two possible interpretations, viz. $\models^{\mathbf{1}}$ and $\models^{\mathbf{0}}$, for schemas from L in frames from $\mathbf{F}$ we have to be precise on the meaning of these correspondences.

3.9. Definition. For $\mathbf{b} \in$ bool we define the schema φ to correspond to the first-order formula P given the interpretation $\models^{\mathbf{b}}$ iff $\forall \mathrm{F} \in \mathbf{F}(\mathrm{F} \models^{\mathbf{b}} \varphi \Leftrightarrow \mathrm{F} \models^{\mathrm{fo}} P)$.

3.10. Proposition. *For atomic actions $a \in \mathrm{At}$, the following correspondences hold in the class $\mathbf{F}$ of frames for $\mathbf{M}$ both for $\mathbf{b} = \mathbf{1}$ and $\mathbf{b} = \mathbf{0}$. The left-hand side of these correspondences is to be understood as a schema in $i \in \mathrm{A}, \varphi \in \mathrm{L}$.*

1. $\mathbf{K}_i[\mathrm{do}_i(a)]\varphi \to [\mathrm{do}_i(a)]\mathbf{K}_i\varphi \sim^{\mathbf{b}}$
   $\forall s_0 \in \mathrm{S} \forall s_1 \in \mathrm{S}(\exists s_2 \in \mathrm{S}(s_2 = \mathtt{r}_0(i,a)(s_0) \,\&\, (s_2, s_1) \in \mathrm{R}(i)) \Rightarrow$
   $\exists s_3 \in \mathrm{S}((s_0, s_3) \in \mathrm{R}(i) \,\&\, s_1 = \mathtt{r}_0(i,a)(s_3)))$
2. $\langle \mathrm{do}_i(a)\rangle\varphi \to [\mathrm{do}_i(a)]\varphi \sim^{\mathbf{b}}$
   $\forall s \in \mathrm{S} \forall s' \in \mathrm{S} \forall s'' \in \mathrm{S}(\mathtt{r}_0(i,a)(s) = s' \,\&\, \mathtt{r}_0(i,a)(s) = s'' \Rightarrow s' = s'')$
3. $[\mathrm{do}_i(a;a)]\varphi \leftrightarrow [\mathrm{do}_i(a)]\varphi \sim^{\mathbf{b}} \forall s \in \mathrm{S}(\mathtt{r}_0(i,a)(\mathtt{r}_0(i,a)(s)) = \mathtt{r}_0(i,a)(s))$
4. $\langle \mathrm{do}_i(a)\rangle\top \sim^{\mathbf{b}} \forall s \in \mathrm{S}(\mathtt{r}_0(i,a)(s) \neq \emptyset)$
5. $\mathbf{A}_i a \to \langle \mathrm{do}_i(a)\rangle\top \sim^{\mathbf{b}} \forall s \in \mathrm{S}(\mathtt{c}_0(i,a)(s) = \mathbf{1} \Rightarrow \mathtt{r}_0(i,a)(s) \neq \emptyset)$

Since the functions $\mathtt{r}_0$ and $\mathtt{c}_0$ are defined for atomic actions only, and the functions $\mathtt{r}^{\mathbf{b}}$ and $\mathtt{c}^{\mathbf{b}}$ — which are the extensions of $\mathtt{r}_0$ and $\mathtt{c}_0$ for arbitrary actions — are constructed out of $\mathtt{r}_0$ and $\mathtt{c}_0$ and have no existence on their own, it is not possible to prove correspondences like those of Proposition 3.10 for non-atomic actions. There simply is no semantic entity to correspond the syntactic schemas with. This implies that it is in general not possible to ensure that arbitrary actions satisfy a certain property. However, it turns out that some of the properties considered above straightforwardly extend from the atomic level to the level of arbitrary actions, regardless of the interpretation that is used. This is in particular the case for the properties of A-realisability and determinism.

3.11. Proposition. *The following lifting results hold for all $\mathrm{F} \in \mathbf{F}$ and $\mathbf{b} \in$ bool:*

- $\forall a \in \mathrm{At}(a$ *is A-realisable for* $\models^{\mathbf{b}}$ *in* $\mathrm{F}) \Rightarrow \forall \alpha \in \mathrm{Ac}(\alpha$ *is A-realisable for* $\models^{\mathbf{b}}$ *in* $\mathrm{F})$
- $\forall a \in \mathrm{At}(a$ *is deterministic for* $\models^{\mathbf{b}}$ *in* $\mathrm{F}) \Rightarrow \forall \alpha \in \mathrm{Ac}(\alpha$ *is deterministic for* $\models^{\mathbf{b}}$ *in* $\mathrm{F})$

Since the range of the function $\mathtt{r}_0$ is the set $\mathrm{S}^{\cdot}$, it follows directly that atomic actions are deterministic in $\mathbf{F}$: for if $a \in \mathrm{At}$, $i \in \mathrm{A}$ and s a state in some model, then $\mathtt{r}_0(i,a)(s)$ is either the empty set, or a single state from S, and hence the frame condition for determinism as given in Proposition 3.10 is satisfied. Using the lifting result obtained in Proposition 3.11 one may then conclude that all actions are deterministic in $\mathbf{F}$.

3.12. Corollary. *All actions $\alpha \in \mathrm{Ac}$ are deterministic in $\mathbf{F}$, both for $\models^{\mathbf{1}}$ and $\models^{\mathbf{0}}$.*

Thus two of the properties formalised in Definition 2.18 can be ensured to hold for arbitrary actions by imposing suitable constraints on the frames for $\mathbf{M}$. For the other

three properties, viz. accordance, idempotence and realisability, constraining the function $\mathtt{r}_0$ for atomic actions does not suffice, since this does not conservatively extend to the class of all actions. That realisability may not be lifted is easily seen by considering the action `fail`. Independent of the realisability of atomic actions, `fail` will never be realisable: the formula $\neg\langle \mathrm{do}_i(\mathtt{fail})\rangle\top$ is valid, both for $\models^{\mathbf{1}}$ and for $\models^{\mathbf{0}}$. The following examples show why accordance and idempotence are in general not to be lifted.

3.13. Example. Consider the language $\mathrm{L}(\Pi, \mathrm{A}, \mathrm{At})$ with $\Pi = \{p, q\}$, $i \in \mathrm{A}$ and At arbitrary. Let $\mathrm{F} \in \mathbf{F}$ be a frame such that the set S of states in F contains at least two elements, say s and t, on which the relation $\mathrm{R}(i)$ is defined to be universal, and the first-order property corresponding with accordance of atomic actions is met. Let π be a valuation such that $\pi(p, s) = \pi(q, s) = \mathbf{1}, \pi(p, t) = \pi(q, t) = \mathbf{0}$. Then we have that $(\mathrm{F}, \pi), s \models^{\mathbf{b}} \mathbf{K}_i[\mathrm{do}_i(\mathtt{confirm}\, p)]q$, and furthermore that $(\mathrm{F}, \pi), s \not\models^{\mathbf{b}} [\mathrm{do}_i(\mathtt{confirm}\, p)]\mathbf{K}_i q$. Hence $\mathrm{F} \not\models^{\mathbf{b}} \mathbf{K}_i[\mathrm{do}_i(\mathtt{confirm}\, p)]q \to [\mathrm{do}_i(\mathtt{confirm}\, p)]\mathbf{K}_i q$, which provides a counterexample to the lifting of accordance.

3.14. Example. Consider the language $\mathrm{L}(\Pi, \mathrm{A}, \mathrm{At})$ with $\Pi = \{p\}$, $i \in \mathrm{A}$ and $\mathrm{At} \supseteq \{a_1, a_2\}$. Consider the frame $\mathrm{F} = \langle \mathrm{S}, \mathrm{R}, \mathtt{r}_0, \mathtt{c}_0\rangle$, where

- $\mathrm{S} = \{s_1, s_2, s_3, s_4\}$
- $\mathrm{R}(i)$ is an arbitrary equivalence relation on S
- $\mathtt{r}_0(i, a_1)(s_1) = s_1 \quad \mathtt{r}_0(i, a_1)(s_2) = s_3 \quad \mathtt{r}_0(i, a_1)(s_3) = s_3 \quad \mathtt{r}_0(i, a_1)(s_4) = \emptyset$
  $\mathtt{r}_0(i, a_2)(s_1) = s_2 \quad \mathtt{r}_0(i, a_2)(s_2) = s_2 \quad \mathtt{r}_0(i, a_2)(s_3) = s_4 \quad \mathtt{r}_0(i, a_2)(s_4) = s_4$
- $\mathtt{c}_0 : \mathrm{A} \times \mathrm{S} \to \mathrm{bool}$ is arbitrary

It is easily checked that both a_1 and a_2 are idempotent in F. However, it is not the case that all actions that can be built on At are idempotent in F. For it holds for arbitrary $\mathbf{b} \in \mathrm{bool}$ that $\mathrm{F} \not\models^{\mathbf{b}} [\mathrm{do}_i((a_1; a_2); (a_1; a_2))]p \leftrightarrow [\mathrm{do}_i(a_1; a_2)]p$. To see this take $\mathrm{M} = (\mathrm{F}, \pi)$ where $\pi(p, s_2) \neq \pi(p, s_4)$. In this model it holds that $\mathrm{M}, s_1 \models^{\mathbf{b}} [\mathrm{do}_i((a_1; a_2); (a_1; a_2))]p \leftrightarrow [\mathrm{do}_i(a_1; a_2)]\neg p$. Hence $\mathrm{M} \not\models^{\mathbf{b}} [\mathrm{do}_i((a_1; a_2); (a_1; a_2))]p \leftrightarrow [\mathrm{do}_i(a_1; a_2)]p$, and therefore also $\mathrm{F} \not\models^{\mathbf{b}} [\mathrm{do}_i((a_1; a_2); (a_1; a_2))]p \leftrightarrow [\mathrm{do}_i(a_1; a_2)]p$. Thus neither for $\models^{\mathbf{1}}$ nor for $\models^{\mathbf{0}}$ is $a_1; a_2$ idempotent in F.

Although we showed in Example 3.13 that accordance is not to be lifted from atomic actions to general ones, we can prove a restricted form of lifting for accordance. That is, if we leave confirmations out of consideration, we can prove that accordance is lifted.

3.15. Proposition. *Let Ac^- be the confirmation-free fragment of Ac, i.e. the fragment built from atomic actions through sequential, conditional or repetitive composition. Then we have for all $\mathrm{F} \in \mathbf{F}$ and for all $\mathbf{b} \in \mathrm{bool}$:*

$$\forall a \in \mathrm{At}(a \textit{ is accordant for } \models^{\mathbf{b}} \textit{ in } \mathrm{F}) \Rightarrow \forall \alpha \in \mathrm{Ac}^-(\alpha \textit{ is accordant for } \models^{\mathbf{b}} \textit{ in } \mathrm{F})$$

The properties of idempotence and (A-)realisability are in general undesirable ones. If all actions were idempotent, it would be impossible to walk the roads by taking one step at a time. Realisability would render the notion of opportunity meaningless and A-realisability would tie ability and opportunity in a way that we feel is unacceptable. Therefore we consider neither the lifting result for A-realisability to be very important, nor the absence of such a result for idempotence and realisability. And even though the property of accordance is, or may be, important, it is not one that typically holds in the lively world of human agents. Therefore we consider this property to be an exceptional one, that holds for selected actions only. Hence also for accordance the absence of a lifting result is not taken too seriously.

3.3 Correctness and feasibility of actions: practical possibility

Within the KARO-architecture, several notions concerning agency may be formalised that are interesting not only from a philosophical point of view, but also when analysing agents in planning systems. The most important one of these notions formalises the knowledge that agents have on their practical possibilities. We consider the notion of practical possibility as pertaining to a pair, consisting of an action and a proposition: agents may have the practical possibility to bring about (truth of) the proposition by performing the action. We think of practical possibility as consisting of two parts, viz. correctness and feasibility. Correctness implies that no external factors will prevent the agent from performing the action and thereby making the proposition true. As such, correctness is defined in terms of opportunity and result: an action is correct for some agent to bring about some proposition iff the agent has the opportunity to perform the action in such a way that its performance results in the proposition being true. Feasibility captures the internal aspect of practical possibility. It states that it is within the agent's capacities to perform the action, and as such is nothing but a reformulation of ability. Together, correctness and feasibility constitute practical possibility.

3.16. Definition. For $\alpha \in \mathrm{Ac}$, $i \in \mathrm{A}$ and $\varphi \in \mathrm{L}$ we define:

- $\mathbf{Correct}_i(\alpha, \varphi) \triangleq \langle \mathrm{do}_i(\alpha) \rangle \varphi$
- $\mathbf{Feasible}_i \alpha \triangleq \mathbf{A}_i \alpha$
- $\mathbf{PracPoss}_i(\alpha, \varphi) \triangleq \mathbf{Correct}_i(\alpha, \varphi) \wedge \mathbf{Feasible}_i \alpha$

The counterintuitive situations that occurred with respect to the ability of agents as described on page 30 do not take root for practical possibility. That is, a lion that has the ability but not the opportunity to eat a zebra will neither have the practical possibility to eat a zebra first and thereafter fly to the moon nor will it have the practical possibility to eat a zebra and rest on its laurels afterwards. Thus even though the

notion of ability suffers from problems like these, the more important notion of practical possibility does not. The importance of practical possibility manifests itself particularly when ascribing — from the outside — certain qualities to an agent. It seems that for the agent itself practical possibilities are relevant in so far as the agent has knowledge on these possibilities. For one may not expect an agent to act on its practical possibilities, like for instance adopt some action as its plan, if the agent does not know of this possibilities. To formalise this kind of knowledge, we introduce the Can-predicate and the Cannot-predicate. The first of these predicates concerns the knowledge of agents on their practical possibilities, the latter predicate does the same for their practical impossibilities.

3.17. Definition. For $\alpha \in \mathrm{Ac}$, $i \in \mathrm{A}$ and $\varphi \in \mathrm{L}$ we define:

- $\mathbf{Can}_i(\alpha, \varphi) \triangleq \mathbf{K}_i\mathbf{PracPoss}_i(\alpha, \varphi)$
- $\mathbf{Cannot}_i(\alpha, \varphi) \triangleq \mathbf{K}_i\neg\mathbf{PracPoss}_i(\alpha, \varphi)$

The Can-predicate and the Cannot-predicate integrate knowledge, ability, opportunity and result, and seem to formalise one of the most important notions of agency. In fact it is probably not too bold to say that knowledge like that formalised through the Can-predicate, although perhaps in a weaker form by taking aspects of uncertainty into account, underlies all acts performed by rational agents, including humans. For rational agents act only if they have some information on both the possibility to perform the act, and its possible outcome. It therefore seems worthwhile to take a closer look at both the Can-predicate and the Cannot-predicate. The following proposition focuses on the behaviour of the *means*-part of the predicates, which is the α in $\mathbf{Can}_i(\alpha, \varphi)$ and $\mathbf{Cannot}_i(\alpha, \varphi)$.

3.18. Proposition. *For all* $\mathbf{b} \in \mathsf{bool}$, $i \in \mathrm{A}$, $\alpha, \alpha_1, \alpha_2 \in \mathrm{Ac}$ *and* $\varphi, \psi \in \mathrm{L}$ *we have:*

1. $\models^{\mathbf{b}} \mathbf{Can}_i(\mathtt{confirm}\,\varphi, \psi) \leftrightarrow \mathbf{K}_i(\varphi \wedge \psi)$
2. $\models^{\mathbf{b}} \mathbf{Cannot}_i(\mathtt{confirm}\,\varphi, \psi) \leftrightarrow \mathbf{K}_i(\neg\varphi \vee \neg\psi)$
3. $\models^{\mathbf{b}} \mathbf{Can}_i(\alpha_1; \alpha_2, \varphi) \leftrightarrow \mathbf{Can}_i(\alpha_1, \mathbf{PracPoss}_i(\alpha_2, \varphi))$
4. $\models^{\mathbf{b}} \mathbf{Can}_i(\alpha_1; \alpha_2, \varphi) \rightarrow \langle \mathrm{do}_i(\alpha_1)\rangle\mathbf{Can}_i(\alpha_2, \varphi)$ *for* α_1 *accordant in* $\mathbf{F}$
5. $\models^{\mathbf{b}} \mathbf{Cannot}_i(\alpha_1; \alpha_2, \varphi) \leftrightarrow \mathbf{Cannot}_i(\alpha_1, \mathbf{PracPoss}_i(\alpha_2, \varphi))$
6. $\models^{\mathbf{b}} \mathbf{Can}_i(\mathtt{if}\,\varphi\,\mathtt{then}\,\alpha_1\,\mathtt{else}\,\alpha_2\,\mathtt{fi}, \psi) \wedge \mathbf{K}_i\varphi \leftrightarrow \mathbf{Can}_i(\alpha_1, \psi) \wedge \mathbf{K}_i\varphi$
7. $\models^{\mathbf{b}} \mathbf{Can}_i(\mathtt{if}\,\varphi\,\mathtt{then}\,\alpha_1\,\mathtt{else}\,\alpha_2\,\mathtt{fi}, \psi) \wedge \mathbf{K}_i\neg\varphi \leftrightarrow \mathbf{Can}_i(\alpha_2, \psi) \wedge \mathbf{K}_i\neg\varphi$
8. $\models^{\mathbf{b}} \mathbf{Cannot}_i(\mathtt{if}\,\varphi\,\mathtt{then}\,\alpha_1\,\mathtt{else}\,\alpha_2\,\mathtt{fi}, \psi) \wedge \mathbf{K}_i\varphi \leftrightarrow \mathbf{Cannot}_i(\alpha_1, \psi) \wedge \mathbf{K}_i\varphi$
9. $\models^{\mathbf{b}} \mathbf{Cannot}_i(\mathtt{if}\,\varphi\,\mathtt{then}\,\alpha_1\,\mathtt{else}\,\alpha_2\,\mathtt{fi}, \psi) \wedge \mathbf{K}_i\neg\varphi \leftrightarrow \mathbf{Cannot}_i(\alpha_2, \psi) \wedge \mathbf{K}_i\neg\varphi$
10. $\models^{\mathbf{b}} \mathbf{Can}_i(\mathtt{while}\,\varphi\,\mathtt{do}\,\alpha\,\mathtt{od}, \psi) \wedge \mathbf{K}_i\varphi \leftrightarrow \mathbf{Can}_i(\alpha, \mathbf{PracPoss}_i(\mathtt{while}\,\varphi\,\mathtt{do}\,\alpha\,\mathtt{od}, \psi)) \wedge \mathbf{K}_i\varphi$
11. $\models^{\mathbf{b}} \mathbf{Can}_i(\mathtt{while}\,\varphi\,\mathtt{do}\,\alpha\,\mathtt{od}, \psi) \wedge \mathbf{K}_i\varphi \rightarrow \langle \mathrm{do}_i(\alpha)\rangle\mathbf{Can}_i(\mathtt{while}\,\varphi\,\mathtt{do}\,\alpha\,\mathtt{od}, \psi)$ *for* α *accordant in* $\mathbf{F}$
12. $\models^{\mathbf{b}} \mathbf{Can}_i(\mathtt{while}\,\varphi\,\mathtt{do}\,\alpha\,\mathtt{od}, \psi) \rightarrow \mathbf{K}_i(\varphi \vee \psi)$

13. $\models^{\mathbf{b}} \mathbf{Cannot}_i(\texttt{while}\,\varphi\,\texttt{do}\,\alpha\,\texttt{od}, \psi) \wedge \mathbf{K}_i\neg\varphi \leftrightarrow \mathbf{K}_i(\neg\varphi \wedge \neg\psi)$
14. $\models^{\mathbf{b}} \mathbf{Cannot}_i(\texttt{while}\,\varphi\,\texttt{do}\,\alpha\,\texttt{od}, \psi) \wedge \mathbf{K}_i\varphi \leftrightarrow \mathbf{Cannot}_i(\alpha; \texttt{while}\,\varphi\,\texttt{do}\,\alpha\,\texttt{od}, \psi) \wedge \mathbf{K}_i\varphi$

Proposition 3.18 does not only serve to support the appropriateness of the Can-predicate and Cannot-predicate as formalising knowledge of practical possibilities, but furthermore provides additional proof that agents are indeed rational. In particular items 6 through 9 and item 14 are genuine indications of the rationality of the agents that we formalised. Consider for example item 7. This item states that whenever an agent knows both that it has the practical possibility to bring about ψ by performing $\texttt{if}\,\varphi\,\texttt{then}\,\alpha_1\,\texttt{else}\,\alpha_2\,\texttt{fi}$ and that the negation of the condition of $\texttt{if}\,\varphi\,\texttt{then}\,\alpha_1\,\texttt{else}\,\alpha_2\,\texttt{fi}$ holds, it also knows that performing the else-part of the conditional composition provides the practical possibility to achieve ψ. Conversely, if agent i knows that it has the practical possibility to bring about ψ by performing α_2 while at the same time knowing that the proposition φ is false, then the agent knows that performing a conditional composition $\texttt{if}\,\varphi\,\texttt{then}\,\alpha_1\,\texttt{else}\,\alpha_2\,\texttt{fi}$ would also bring about ψ, regardless of α_1. For since it knows that $\neg\varphi$ holds, it knows that this compositional composition comes down to the else-part α_2. Items 4 and 11 explicitly use the accordance of actions. For it is exactly this property of accordance that causes the agent's knowledge on its practical possibilities to persist under execution of the first part of the sequential composition in item 4 and the body of the while-loop in item 11.

In the following proposition we characterise the relation between the Can-predicate and the Cannot-predicate. Furthermore some properties are presented that concern the *end*-part of these predicates, i.e. the φ in $\mathbf{Can}_i(\alpha, \varphi)$ and $\mathbf{Cannot}_i(\alpha, \varphi)$.

3.19. Proposition. *For all* $\mathbf{b} \in \mathrm{bool}$, $i \in \mathrm{A}$, $\alpha \in \mathrm{Ac}$ *and* $\varphi, \psi \in \mathrm{L}$ *we have:*

1. $\models^{\mathbf{b}} \mathbf{Can}_i(\alpha, \varphi) \rightarrow \neg\mathbf{Can}_i(\alpha, \neg\varphi)$
2. $\models^{\mathbf{b}} \mathbf{Can}_i(\alpha, \varphi) \rightarrow \neg\mathbf{Cannot}_i(\alpha, \varphi)$
3. $\models^{\mathbf{b}} \mathbf{Can}_i(\alpha, \varphi) \rightarrow \mathbf{Cannot}_i(\alpha, \neg\varphi)$
4. $\models^{\mathbf{b}} \mathbf{Can}_i(\alpha, \varphi \wedge \psi) \leftrightarrow \mathbf{Can}_i(\alpha, \varphi) \wedge \mathbf{Can}_i(\alpha, \psi)$
5. $\models^{\mathbf{b}} \mathbf{Cannot}_i(\alpha, \varphi) \vee \mathbf{Cannot}_i(\alpha, \psi) \rightarrow \mathbf{Cannot}_i(\alpha, \varphi \wedge \psi)$
6. $\models^{\mathbf{b}} \mathbf{Can}_i(\alpha, \varphi) \vee \mathbf{Can}_i(\alpha, \psi) \rightarrow \mathbf{Can}_i(\alpha, \varphi \vee \psi)$
7. $\models^{\mathbf{b}} \mathbf{Cannot}_i(\alpha, \varphi \vee \psi) \leftrightarrow \mathbf{Cannot}_i(\alpha, \varphi) \wedge \mathbf{Cannot}_i(\alpha, \psi)$
8. $\models^{\mathbf{b}} \mathbf{Can}_i(\alpha, \varphi) \wedge \mathbf{K}_i[\mathrm{do}_i(\alpha)](\varphi \rightarrow \psi) \rightarrow \mathbf{Can}_i(\alpha, \psi)$
9. $\models^{\mathbf{b}} \mathbf{Cannot}_i(\alpha, \varphi) \wedge \mathbf{K}_i[\mathrm{do}_i(\alpha)](\psi \rightarrow \varphi) \rightarrow \mathbf{Cannot}_i(\alpha, \psi)$

Even more than Proposition 3.18 does Proposition 3.19 make out a case for the rationality of agents. Take for example item 3, which states that whenever an agent knows that it has the practical possibility to achieve φ by performing α it also knows that α does not provide for a means to achieve $\neg\varphi$. Items 4 through 7 deal with the

decomposition of the end-part of the Can-predicate and the Cannot-predicate, which behaves as desired. Note that the reverse implication of item 5 is not valid: it is quite possible that even though an agent knows that α is not correct to bring about $\varphi \wedge \psi$ it might still be that it knows that α is correct for either φ or ψ. An analogous line of reasoning shows the invalidity of the reverse implication of item 6. Items 8 and 9 formalise that agents can extend their knowledge on their practical (im)possibilities by combining it with their knowledge on the (conditional) results of actions.

3.4 Proof theory

Here we present a proof theory for the semantic framework defined in the previous section. In general the purpose of a proof theory is to provide a syntactic counterpart of the semantic notion of validity for a given interpretation and a given class of models. The idea is to define a predicate denoting deducibility, which holds for a given formula iff the formula is valid. This predicate is to be defined purely syntactical, i.e. it should depend only on the syntactic structure of formulae, without making any reference to semantic notions as truth, validity, satisfiability etc. We present two such predicates, viz. $\vdash^{\mathbf{1}}$ and $\vdash^{\mathbf{0}}$, which characterise the notions of validity associated with $\models^{\mathbf{1}}$ and $\models^{\mathbf{0}}$, respectively. The definition of these predicates is based on a set of axioms and proof rules, which together constitute a proof system. The proof systems that we define deviate somewhat from the ones that are common in (modal) logics, the most notable difference being the use of infinitary proof rules. Given the relative rarity of this kind of rules, we feel that some explanation is justified.

3.4.1 Infinitary proof rules

The proof rules that are commonly employed in proof systems, are sentences of the form $\mathrm{P}_1, \ldots, \mathrm{P}_m \,/\, \mathrm{C}$, where the premises $\mathrm{P}_1, \ldots, \mathrm{P}_m$ and the conclusion C are elements of the language under consideration. Informally, a rule like this denotes that one may deduce C as soon as $\mathrm{P}_1, \ldots, \mathrm{P}_m$ have been deduced. An infinitary[3] proof rule is a rule containing an infinite number of premises. Although not very common, infinitary proof rules have been used in a number of proof systems: Hilbert used an infinitary proof rule in axiomatising number theory [47], Schütte uses infinitary proof rules in a number of systems [117], Gabbay proposes the use of an infinitary rule to characterise the notion of irreflexivity prooftheoretically [34], and both Kröger [73] and Goldblatt [41, 42] use infinitary proof rules in logics of action.

[3] We decided to follow the terminology of Goldblatt [41, 42] and refer to these rules as being infinitary. Other authors call these rules infinite [73, 117].

In finitary proof systems proofs can be carried out completely within the formal system. A proof is usually taken to be a finite-length sequence of formulae that are either axioms of the proof system or conclusions of proof rules applied to formulae that appear earlier in the sequence. Since finitary proof rules can be applied as soon as all of their finitely many premises have been deduced, there is no need to step outside of the formal system. In order to apply an infinitary rule, a meta-logical investigation on the deducibility of the (infinitely many) premises needs to be carried out, which makes it in general impossible to carry out proofs completely within the proof system. As such, proofs are no longer 'schematically' constructed, and theorems are not recursively enumerable. However, there is also a number of advantages associated with the use of infinitary proof rules. The first of these is that for some systems *strong completeness* can be achieved using infinitary proof rules, whereas this is not possible using finitary proof rules (cf. [41, 117]). The notion of strong completeness implies that fewer sets of formulae are consistent, and in particular that sets of formulae that are seen to be inconsistent can also be proved to be so. After the presentation of the proof systems, we will return to the property of strong completeness in the presence of infinitary rules. Besides the possibility to achieve strong completeness when using infinitary proof rules, there are two other arguments that influenced our choice to use this kind of rules. The first of these is the intuitive acceptability of this kind of rules. In particular when dealing with notions with an infinitary character, like for instance while-loops, infinitary proof rules provide a much better formalisation of human intuition on the nature of these notions than do finitary proof rules. The second, perhaps less convincing but certainly more compelling, argument is given by the fact that our attempts to come up with finitary axiomatisations remained unavailing.

3.4.2 Logics of capabilities

Before presenting the actual axiomatisations, we first make some notions precise that were already informally discussed above. An axiom is a schema in L. A proof rule is a sentence of the form $\varphi_1, \varphi_2, \ldots / \psi$ where $\varphi_1, \varphi_2, \ldots, \psi$ are schemas in L. A proof system is a pair consisting of a set of axioms and a set of proof rules. As mentioned above, the presence of infinitary proof rules forces us to adopt a more abstract approach to the notions of deducibility and theorem than the one commonly employed in finitary proof systems. Usually, a formula φ is defined to be a theorem of some proof system if there exist a finite-length sequence of formulae of which φ is the last element and such that each formula in the sequence is either an instance of an axiom or the conclusion of a proof rule applied to earlier members of the sequence. An alternative formulation, which is equally usable in finitary and in infinitary proof systems, is to define φ to be a theorem of a proof system iff it belongs to the smallest subset of L containing all (instances of all)

axioms and closed under the proof rules. This latter notion of deducibility is actually the one that we will employ here. We define a logic for a given proof system to be a subset of L containing all instances of the axioms of the proof system and closed under its proof rules. A formula is a theorem for a given proof system iff it is an element of the smallest logic for the proof system. These notions are formalised in Definitions 3.23 through 3.25.

To axiomatise the behaviour of while-loops we propose two infinitary rules. Both these rules are based on the idea to equate an action $\mathtt{while}\,\varphi\,\mathtt{do}\,\alpha\,\mathtt{od}$ with the infinite set $\mathrm{CS}(\mathtt{while}\,\varphi\,\mathtt{do}\,\alpha\,\mathtt{od})$ of finite computation sequences. The two proof rules take as their premises an infinite set of formulae built around the set $\mathrm{CS}(\mathtt{while}\,\varphi\,\mathtt{do}\,\alpha\,\mathtt{od})$ and have as their conclusion a formula built around $\mathtt{while}\,\varphi\,\mathtt{do}\,\alpha\,\mathtt{od}$. To make this idea of 'building formulae around actions' explicit, we introduce the concept of *admissible forms*. The notion of admissible forms as given in Definition 3.20 is an extension of that used by Goldblatt in his language of program schemata [41]. In his investigation of infinitary proof rules, Kröger found that, in order to prove completeness, he needed rules in which the context of the while-loop and its set of computation sequences is taken into account [73]. The concept of admissible forms provides an abstract generalisation of this idea of taking contexts into account.

3.20. Definition. The set of admissible forms for L, denoted by $\mathrm{Afm}(\mathrm{L})$, is defined by the following BNF.

$$\phi ::= \# \mid [\mathrm{do}_i(\alpha)]\phi \mid \mathbf{K}_i\phi \mid \psi \to \phi$$

where $i \in \mathrm{A}$, $\alpha \in \mathrm{Ac}$ and $\psi \in \mathrm{L}$. We use the letter ϕ as a typical element of $\mathrm{Afm}(\mathrm{L})$.

Usually 'admissible form' is abbreviated to 'afm'. By definition, each afm has a unique occurrence of the special symbol $\#$. By instantiating this symbol with a formula from L, afms are turned into genuine formula. If ϕ is an afm and $\psi \in \mathrm{L}$ is some formula we denote by $\phi(\psi)$ the formula that is obtained by replacing (the unique occurrence of) $\#$ in ϕ by ψ.

The following definition introduces two abbreviations that will be used in formulating the infinitary rules.

3.21. Definition. For all $\psi, \varphi \in \mathrm{L}$, $i \in \mathrm{A}$, $\alpha \in \mathrm{Ac}$ and $l \in \mathbb{N}$ we define:

- $\psi_l(i, \varphi, \alpha) \triangleq [\mathrm{do}_i((\mathtt{confirm}\,\varphi; \alpha)^l; \mathtt{confirm}\,\neg\varphi)]\psi$
- $\varphi_l(i, \alpha) \triangleq \mathbf{A}_i((\mathtt{confirm}\,\varphi; \alpha)^l; \mathtt{confirm}\,\neg\varphi)$

The formulae introduced in Definition 3.21 are used to define the premises of the infinitary rules. The rule formalising the behaviour of while-loops with respect to results

and opportunities has an infinite set of premises of the form $\phi(\psi_l(i,\varphi,\alpha))$ with $l \in \mathbb{N}$ and $\phi \in \mathrm{Afm(L)}$. The conclusion of this rule is the formula $\phi([\mathrm{do}_i(\mathtt{while}\,\varphi\,\mathtt{do}\,\alpha\,\mathtt{od})]\psi)$. Leaving the context provided by ϕ out of consideration, this rules intuitively states that if it is deducible that ψ holds after executing any of the finite computation sequences that possibly constitute $\mathtt{while}\,\varphi\,\mathtt{do}\,\alpha\,\mathtt{od}$, it is also deducible that ψ holds after executing $\mathtt{while}\,\varphi\,\mathtt{do}\,\alpha\,\mathtt{od}$. The rule used in formalising the ability of agents for while-loops has as its premises the set $\phi(\neg(\varphi_l(i,\alpha)))$ for $l \in \mathbb{N}, \phi \in \mathrm{Afm(L)}$, and a conclusion $\phi(\neg\mathbf{A}_i\mathtt{while}\,\varphi\,\mathtt{do}\,\alpha\,\mathtt{od})$. This rule states that whenever it is deducible that an agent i is not capable of performing any of the finite computation sequences that syntactically constitute the while-loop, it is also deducible that the agent is incapable of performing the while-loop itself. Or read in its contrapositive form, that an agent is able to perform a while-loop only if it is able to perform some finite-length sequence of confirmations and actions constituting the while-loop. As such, this rule is easily seen to be the proof-theoretic counterpart of the negated version of the (semantic) definition of $\mathsf{c}^{\mathbf{b}}$ for while-loops. For read in its negative form this semantic definition states that $\mathsf{c}^{\mathbf{b}}(i,\mathtt{while}\,\varphi\,\mathtt{do}\,\alpha\,\mathtt{od})(s) = \mathbf{0}$ iff $\mathsf{c}^{\mathbf{b}}(i,\varphi_l(i,\alpha))(s) = \mathbf{0}$ for all $l \in \mathbb{N}$.

The axioms that are used to build the two proof systems are formulated using the necessity operator for actions, i.e. $[\mathrm{do}\_(\_)]\_$, rather than its dual $\langle\mathrm{do}\_(\_)\rangle\_$. The reason for this is essentially one of convenience: in proving completeness of the axiomatisations it turns out to be useful to deal with two necessity operators, viz. $\mathbf{K}\_$ and $[\mathrm{do}\_(\_)]\_$, to allow proofs by analogy. Since $[\mathrm{do}\_(\_)]\_$ and $\langle\mathrm{do}\_(\_)\rangle\_$ are interdefinable this does not make up for essential differences.

3.22. Definition. The following axioms and proof rules are used to constitute the two proof systems that we consider here. Both the axioms as well as the premises and conclusions of the proof rules are to be taken as schemas in $i \in \mathrm{A}, \varphi, \psi \in \mathrm{L}$ and $\alpha, \alpha_1, \alpha_2 \in \mathrm{Ac}$. The ϕ occurring in the two infinitary rules ΩI and ΩIA is taken to be a meta-variable ranging over $\mathrm{Afm(L)}$.

A1. All propositional tautologies and their epistemic and dynamic instances
A2. $\mathbf{K}_i(\varphi \to \psi) \to (\mathbf{K}_i\varphi \to \mathbf{K}_i\psi)$
A3. $\mathbf{K}_i\varphi \to \varphi$
A4. $\mathbf{K}_i\varphi \to \mathbf{K}_i\mathbf{K}_i\varphi$
A5. $\neg\mathbf{K}_i\varphi \to \mathbf{K}_i\neg\mathbf{K}_i\varphi$
A6. $[\mathrm{do}_i(\alpha)](\varphi \to \psi) \to ([\mathrm{do}_i(\alpha)]\varphi \to [\mathrm{do}_i(\alpha)]\psi)$
A7. $[\mathrm{do}_i(\mathtt{confirm}\,\varphi)]\psi \leftrightarrow (\neg\varphi \vee \psi)$
A8. $[\mathrm{do}_i(\alpha_1;\alpha_2)]\varphi \leftrightarrow [\mathrm{do}_i(\alpha_1)][\mathrm{do}_i(\alpha_2)]\varphi$
A9. $[\mathrm{do}_i(\mathtt{if}\,\varphi\,\mathtt{then}\,\alpha_1\,\mathtt{else}\,\alpha_2\,\mathtt{fi})]\psi \leftrightarrow$
$([\mathrm{do}_i(\mathtt{confirm}\,\varphi;\alpha_1)]\psi \wedge [\mathrm{do}_i(\mathtt{confirm}\,\neg\varphi;\alpha_2)]\psi)$
A10. $[\mathrm{do}_i(\mathtt{while}\,\varphi\,\mathtt{do}\,\alpha\,\mathtt{od})]\psi \leftrightarrow ([\mathrm{do}_i(\mathtt{confirm}\,\neg\varphi)]\psi\wedge$

$[\mathrm{do}_i(\mathtt{confirm}\,\varphi;\alpha)][\mathrm{do}_i(\mathtt{while}\,\varphi\,\mathtt{do}\,\alpha\,\mathtt{od})]\psi)$

A11. $[\mathrm{do}_i(\alpha)]\varphi \vee [\mathrm{do}_i(\alpha)]\neg\varphi$

A12. $\mathbf{A}_i\mathtt{confirm}\,\varphi \leftrightarrow \varphi$

$\mathrm{A13_1}$. $\mathbf{A}_i(\alpha_1;\alpha_2) \leftrightarrow \mathbf{A}_i\alpha_1 \wedge [\mathrm{do}_i(\alpha_1)]\mathbf{A}_i\alpha_2$

$\mathrm{A13_0}$. $\mathbf{A}_i(\alpha_1;\alpha_2) \leftrightarrow \mathbf{A}_i\alpha_1 \wedge \langle\mathrm{do}_i(\alpha_1)\rangle\mathbf{A}_i\alpha_2$

A14. $\mathbf{A}_i\mathtt{if}\,\varphi\,\mathtt{then}\,\alpha_1\,\mathtt{else}\,\alpha_2\,\mathtt{fi} \leftrightarrow$
$(\mathbf{A}_i\mathtt{confirm}\,\varphi;\alpha_1 \vee \mathbf{A}_i\mathtt{confirm}\,\neg\varphi;\alpha_2)$

$\mathrm{A15_1}$. $\mathbf{A}_i\mathtt{while}\,\varphi\,\mathtt{do}\,\alpha\,\mathtt{od} \leftrightarrow (\mathbf{A}_i(\mathtt{confirm}\,\neg\varphi)\vee$
$(\mathbf{A}_i\mathtt{confirm}\,\varphi;\alpha \wedge [\mathrm{do}_i(\mathtt{confirm}\,\varphi;\alpha)]\mathbf{A}_i\mathtt{while}\,\varphi\,\mathtt{do}\,\alpha\,\mathtt{od}))$

$\mathrm{A15_0}$. $\mathbf{A}_i\mathtt{while}\,\varphi\,\mathtt{do}\,\alpha\,\mathtt{od} \leftrightarrow (\mathbf{A}_i(\mathtt{confirm}\,\neg\varphi)\vee$
$(\mathbf{A}_i\mathtt{confirm}\,\varphi;\alpha \wedge \langle\mathrm{do}_i(\mathtt{confirm}\,\varphi;\alpha)\rangle\mathbf{A}_i\mathtt{while}\,\varphi\,\mathtt{do}\,\alpha\,\mathtt{od}))$

R1. $\phi(\psi_l(i,\varphi,\alpha))$ all $l \in \mathbb{N}$ / $\phi([\mathrm{do}_i(\mathtt{while}\,\varphi\,\mathtt{do}\,\alpha\,\mathtt{od})]\psi)$ ΩI

R2. $\phi(\neg(\varphi_l(i,\alpha)))$ all $l \in \mathbb{N}$ / $\phi(\neg\mathbf{A}_i\mathtt{while}\,\varphi\,\mathtt{do}\,\alpha\,\mathtt{od})$ ΩIA

R3. $\varphi,\ \varphi \rightarrow \psi\ /\ \psi$ MP

R4. $\varphi\ /\ \mathbf{K}_i\varphi$ KN

R5. $\varphi\ /\ [\mathrm{do}_i(\alpha)]\varphi$ AN

Most of the axioms are fairly obvious, in particular given the discussion on the validities presented in Section 3.1.3. Rule R1, the Omega Iteration rule, is adopted from the axiomatisations given by Goldblatt [41, 42]. Both ΩI and rule R2, which is the Omega Iteration rule for Ability, were already discussed above. Rule R3 is the rule of Modus Ponens, well known from, and used in, both classical and modal logics. R4 and R5 are both instances of the rule of necessitation, which is known to hold for necessity operators. These rules state that whenever some formula is deducible, it is also deducible that an arbitrary agent knows the formula, and that all events have this formula among their conditional results, respectively. Axioms A2 and A6, and the rules R4 and R5 indicate that both knowledge and conditional results are formalised through normal modal operators.

The axioms and proof rules given above are used to define two different proof systems. One of these proof systems embodies the optimistic view on abilities in the counterfactual state of affairs, the other employs a pessimistic view.

3.23. Definition. The proof system Σ_1 contains the axioms A1 through A12, $\mathrm{A13_1}$, A14, $\mathrm{A15_1}$ and the proof rules R1 through R5. The proof system Σ_0 contains the axioms A1 through A12, $\mathrm{A13_0}$, A14, $\mathrm{A15_0}$ and the proof rules R1 through R5.

As mentioned above, a logic for a given proof system is a set encompassing the proof system.

3.24. Definition. A $\mathbf{b}$-logic is a set Λ that contains all the instances of the axioms of $\Sigma_{\mathbf{b}}$ and is closed under the proof rules of $\Sigma_{\mathbf{b}}$. The intersection of all $\mathbf{b}$-logics, which is itself a $\mathbf{b}$-logic, viz. the smallest one, is denoted by $\mathrm{LCap}_{\mathbf{b}}$. Whenever the underlying proof system is either irrelevant or clear from the context, we refer to a $\mathbf{b}$-logic simply as a logic.

Deducibility in a given proof system is now defined as being an element of the smallest logic for the proof system.

3.25. Definition. For Λ some logic, the unary predicate $\vdash^{\Lambda} \subseteq \mathrm{L}$ is defined by: $\vdash^{\Lambda} \varphi \Leftrightarrow \varphi \in \Lambda$. As an abbreviation we occasionally write $\vdash^{\mathbf{b}} \varphi$ for $\vdash^{\mathrm{LCap}_{\mathbf{b}}}$. Whenever $\vdash^{\Lambda} \varphi$ holds we say that φ is deducible in Λ or alternatively that φ is a theorem of Λ.

The proof systems $\Sigma_{\mathbf{1}}$ and $\Sigma_{\mathbf{0}}$ provide sound and complete axiomatisations of validity for $\models^{\mathbf{1}}$ and $\models^{\mathbf{0}}$ respectively.

3.26. Theorem. *For* $\mathbf{b} \in \mathrm{bool}$ *and all* $\varphi \in \mathrm{L}$ *we have:* $\vdash^{\mathbf{b}} \varphi \Leftrightarrow \models^{\mathbf{b}} \varphi$.

Besides the notion of deducibility *per se*, it is also interesting to look at deducibility from a set of premises. In modal logics one may distinguish two notions of deducibility from premises. In the first of these, the premises are considered to be additional axioms, on which also rules of necessitation may be applied. The second notion of deducibility allows necessitation only on the axioms of the proof system, and not on the premises. This latter notion of deducibility is perhaps the more natural one, and is in fact the one that we will concentrate on. A thorough discussion on the relation between the two notions of deducibility is to be found elsewhere [56, 95].

To account for deducibility from premises with respect to the alternative notion of deducibility as being element of some set of formulae, we introduce the notion of a theory of a logic. Corresponding to the idea that the rules of necessitation are not to be applied on premises, a theory is not demanded to be closed under these rules. A formula is now defined to be deducible from some set of premises iff it is contained in every theory that encompasses the set of premises.

3.27. Definition. For Λ some logic, we define a Λ-theory to be any subset Θ of L that contains Λ and is closed under the rules ΩI, ΩIA, and MP.

3.28. Definition. Let Λ be some logic and $\Phi \cup \{\varphi\} \subseteq \mathrm{L}$. The binary relation $\vdash^{\Lambda} \subseteq \wp(\mathrm{L}) \times \mathrm{L}$ is defined by:

$$\Phi \vdash^{\Lambda} \varphi \Leftrightarrow \varphi \in \bigcap\{\Gamma \subseteq \mathrm{L} \mid \Phi \subseteq \Gamma \text{ and } \Gamma \text{ is a } \Lambda\text{-theory}\}$$

Whenever $\Phi \vdash^{\Lambda} \varphi$ we say that φ is deducible from Φ in Λ. A set $\Phi \subseteq \mathrm{L}$ is called Λ-inconsistent iff $\Phi \vdash^{\Lambda} \bot$, and Λ-consistent iff it is not Λ-inconsistent.

Given the 'overloading' of the symbol $\vdash^{\Lambda}$ as representing both deducibility *per se* and deducibility from premises, it is highly desirable that the two uses of this symbol coincide in the case that the set of premises is empty: deducibility from an empty set of premises should not differ from deducibility *per se*.

3.29. Proposition. *For Λ some logic and $\varphi \in \mathrm{L}$ we have: $\vdash^{\Lambda} \varphi \Leftrightarrow \emptyset \vdash^{\Lambda} \varphi$.*

As already mentioned before, using infinitary rules to describe the behaviour of while-loops allows one to achieve strong completeness, the notion which states that every consistent set of formulae is simultaneously satisfiable. Achieving strong completeness is in general not possible when just finitary rules are used. To see this consider the set $\Omega = \{[\mathrm{do}_i(a^k)]p \mid k \in \mathbb{N}\} \cup \{\langle \mathrm{do}_i(\mathtt{while}\, p\, \mathtt{do}\, a\, \mathtt{od})\rangle \top\}$. It is obvious that Ω is not satisfiable. For whenever $\mathrm{M}, s \models^{\mathbf{b}} [\mathrm{do}_i(a^k)]p$ for all $k \in \mathbb{N}$ then execution of $\mathtt{while}\, p\, \mathtt{do}\, a\, \mathtt{od}$ does not terminate, and hence $\mathrm{M}, s \not\models \langle \mathrm{do}_i(\mathtt{while}\, p\, \mathtt{do}\, a\, \mathtt{od})\rangle \top$. However, when using just finitary rules to describe while-loops (like for instance the well-known Hoare rule [49]), the set Ω will be consistent. For when restricting oneself to finitary rules, consistency of an infinite set of formula corresponds to consistency of each of its finite subsets. And in every axiomatisation that is to be sound, all finite subsets of Ω should be consistent, and therefore Ω itself is consistent. In the infinitary proof systems $\Sigma_{\mathbf{1}}$ and $\Sigma_{\mathbf{0}}$ it holds that $\bot$ is deducible from Ω, i.e. Ω is inconsistent. More in general, the property of strong completeness holds for both $\Sigma_{\mathbf{1}}$ and $\Sigma_{\mathbf{0}}$.

3.30. Proposition. *The proof systems $\Sigma_{\mathbf{1}}$ and $\Sigma_{\mathbf{0}}$ are strongly complete, i.e. every set $\Phi \subseteq \mathrm{L}$ that is $\mathrm{LCap}_{\mathbf{b}}$-consistent is $\models^{\mathbf{b}}$-satisfiable.*

Just as deducibility *per se* is the proof theoretic counterpart of the semantic notion of validity, there is also a semantic counterpart to the notion of deducibility from premises.

3.31. Proposition. *For $\mathbf{b} \in$ bool, $\Phi \subseteq \mathrm{L}$ and $\varphi \in \mathrm{L}$ we have:*

$\Phi \vdash^{\mathbf{b}} \varphi \Leftrightarrow \Phi \models^{\mathbf{b}} \varphi$

where $\Phi \models^{\mathbf{b}} \varphi$ iff $\mathrm{M}, s \models^{\mathbf{b}} \Phi$ implies $\mathrm{M}, s \models^{\mathbf{b}} \varphi$ for all $\mathrm{M} \in \mathbf{M}$ with state s.

In the light of the strong completeness property, Proposition 3.31 is not very surprising. In fact, the left-to-right implication is a direct consequence of the strong completeness property. The right-to-left implication follows from the observation that the set of formulae that is satisfied in some world forms a theory.

3.5 Summary and conclusions

In this chapter we presented the first two of the formal systems considered in this thesis, which are all built on the core framework developed in the previous chapter. As far as the syntax and the models are concerned, these first formal systems equal the core framework. We explained our intuition on the composite behaviour of results, opportunities and abilities, and presented two formal interpretations that comply with this intuition. These interpretations differ in their treatment of abilities of agents for sequentially composed actions. We showed how some of the additional properties of actions that were proposed at the end of the previous chapter can be brought about by imposing suitable constraints on the models. In particular it holds for the properties of determinism and A-realisability that constraints imposed on atomic actions inherit to all actions. Using the various modalities present in the framework, we proposed a formalisation of the knowledge of agents on their practical possibilities, a notion which captures an important aspect of agency, particularly in the context of planning agents. We presented two proof systems that syntactically characterise the notion of validity in the two interpretations that we defined. The most remarkable aspect of these proof systems is the use of infinitary proof rules, which on the one hand allows for a better correspondence between the semantic notion of validity and its syntactic counterpart, on the other hand forces one to generalise the usual notions of proof and theorem.

3.5.1 Possible extensions

In the KARO-architecture we proposed two definitions for the ability of agents to execute a sequentially composed action $\alpha_1;\alpha_2$ in cases where execution of α_1 leads to the counterfactual state of affairs. The simplicity of these definitions, both at a conceptual and at a technical level, may lead to counterintuitive situations. Recall that using the so-called optimistic approach it is possible that an agent is considered to be capable of performing $\alpha;\mathtt{fail}$, whereas in the pessimistic approaches agents are declared unable to perform $\alpha;\mathtt{skip}$, for $\alpha \in \mathrm{Ac}$. A more realistic approach would be not to treat all actions equally, but instead to determine for each action individually whether it makes sense to declare an agent (un)able to perform the action in the counterfactual state of affairs. One way to formalise this consists of extending the models from $\mathbf{M}$ with an additional function $\mathtt{t} : \mathrm{A} \times \mathrm{Ac} \to \mathrm{S} \to \mathrm{S}$ which is such that $\mathtt{t}(i,\alpha)(s) = \mathtt{r}(i,\alpha)(s)$ whenever $\mathtt{r}(i,\alpha)(s) \neq \emptyset$. Hence in the case that $\mathtt{r}(i,\alpha)(s) \neq \emptyset$, $\mathtt{t}(i,\alpha)(s)$ equals $\mathtt{r}(i,\alpha)(s)$ and in other cases $\mathtt{t}(i,\alpha)(s)$ is definitely not empty. The ability for the sequential composition is then defined by

$$\mathtt{c}(i,\alpha_1;\alpha_2)(s) = \mathbf{1} \Leftrightarrow \mathtt{c}(i,\alpha_1)(s) = \mathbf{1} \,\&\, \mathtt{c}(i,\alpha_2)(\mathtt{t}(i,\alpha_1)(s)) = \mathbf{1}$$

Applying this definition would yield that $\mathbf{A}_i\alpha_1;\mathtt{fail}$ is no longer satisfiable, and that $\mathbf{A}_i\alpha_1;\mathtt{skip}$ holds in cases where $\mathbf{A}_i\alpha_1 \wedge \neg\langle \mathrm{do}_i(\alpha_1)\rangle\top$ is true. A special instantiation of this approach corresponds to the idea that abilities of agents do not tend to change. Therefore it could seem reasonable to assume that agents retain their abilities when ending up in the counterfactual state of affairs. Formally this can be brought about by demanding $\mathrm{t}(i,\alpha)(s)$ to equate s in cases where $\mathrm{r}(i,\alpha)(s) = \emptyset$. Since this is but a special case of the general idea discussed above, it also avoids the counterintuitive situations where agents are declared to be able to do $\alpha;\mathtt{fail}$ or unable to do $\alpha;\mathtt{skip}$.

Another possible extension of the framework presented in this chapter concerns the introduction of actions with typical, as opposed to certain, effects. Intuitively, the execution of such an action will typically lead to a particular effect, but unexpected events may prevent this from happening. Actions like these are supposed to be better formalisations of human acts, which in general indeed have uncertain effects. A first proposal to formalise actions with typical effects in the KARO-architecture is due to Dunin-Keplicz & Radzikowska [28].

3.5.2 Bibliographical notes

The formal framework presented in this chapter is a revised and considerably extended version of the one presented in [54], to which some elements of [57] are added. In [54] we presented a genuinely modal version of the framework defined by Moore [99, 100]. Although he combines knowledge and action in a modal way, Moore uses the first-order correspondences to reason with. Without going into detail too much, we would like to point out some characteristic features of Moore's framework, especially when compared to the one we presented above. First of all, the developments in research on modal logic that followed Moore's work, make this first system look a little bit outdated now: the notation seems unnecessarily complex, and the system is altogether rather difficult to understand. By using well-known and established concepts from epistemic and dynamic logic we hope to have overcome this problem. Secondly, whereas we treat abilities in their own right, Moore strives to define ability in terms of knowledge on actions. Hence, whereas we are mainly interested in how the ability to perform a composite action depends on the ability to perform its components, Moore is more interested in the nature of ability. Lastly, since Moore is reasoning with first-order correspondences, he does not bother to provide a (modal) proof theory for his framework, but instead uses classical first-order logic. The proof theory as we present it in Section 3.4 is inspired by the one given by Goldblatt in his account of so called program logic [41]. Together with Kröger [73], Goldblatt is one of the very few to use infinitary proof rules in a logic of action. The motivation for both Kröger and Goldblatt to use infinitary rules is twofold. In addition to having the possibility to achieve strong completeness, they furthermore feel that infinitary actions

like while-loops should be axiomatised using infinitary rules.

Two other influential formalisations of agency in AI that we would like to mention explicitly, are the one of Cohen & Levesque [20] and that of Rao & Georgeff [108, 109, 110]. Both these formalisations are based on modal logics and employ Kripke-style possible worlds models, but that is as far as the similarity goes. Cohen & Levesque's main aim is to provide a formalisation of motivational attitudes which complies with the philosophical criteria proposed by Bratman [9]. For this they use four primary modalities: one representing the beliefs of an agent, one representing the agents' goals, one modality indicating that some action will happen next, and one indicating that some action has just happened. The motivational attitudes of intention and commitment are defined in terms of these primary modalities. Whereas we are interested in informational attitudes and actions in their own right, Cohen & Levesque consider these only as a basis to define motivational attitudes. In our treatment of motivational attitudes (Chapter 6) we elaborate on the system of Cohen & Levesque. The system of Rao & Georgeff is based on three primitive modalities: beliefs, desires, and intentions. Semantically their formalism is based on a branching model of time, in which belief-, desire- and intention-accessible worlds are themselves branching time structures. The emphasis of their approach lies within formalising the revision of intentions, beliefs, and goals. Other systems besides those of Cohen & Levesque and Rao & Georgeff can be found in the overview article of Wooldridge & Jennings [129] and in the proceedings of several workshops on agents [32, 58, 79, 130, 131].

3.6 Selected proofs

3.6. Proposition. *For* $\mathbf{b} \in \mathsf{bool}$, $i \in \mathrm{A}$, $\alpha, \alpha_1, \alpha_2 \in \mathrm{Ac}$ *and* $\varphi \in \mathrm{L}$ *we have:*

1. $\models^{\mathbf{b}} \mathbf{A}_i \mathtt{confirm}\, \varphi \leftrightarrow \varphi$
2. $\models^{\mathbf{1}} \mathbf{A}_i \alpha_1; \alpha_2 \leftrightarrow \mathbf{A}_i \alpha_1 \wedge [\mathrm{do}_i(\alpha_1)] \mathbf{A}_i \alpha_2$
3. $\models^{\mathbf{0}} \mathbf{A}_i \alpha_1; \alpha_2 \leftrightarrow \mathbf{A}_i \alpha_1 \wedge \langle \mathrm{do}_i(\alpha_1) \rangle \mathbf{A}_i \alpha_2$
4. $\models^{\mathbf{b}} \mathbf{A}_i \mathtt{if}\, \varphi\, \mathtt{then}\, \alpha_1\, \mathtt{else}\, \alpha_2\, \mathtt{fi} \leftrightarrow ((\varphi \wedge \mathbf{A}_i \alpha_1) \vee (\neg\varphi \wedge \mathbf{A}_i \alpha_2))$
5. $\models^{\mathbf{1}} \mathbf{A}_i \mathtt{while}\, \varphi\, \mathtt{do}\, \alpha\, \mathtt{od} \leftrightarrow (\neg\varphi \vee (\varphi \wedge \mathbf{A}_i \alpha \wedge [\mathrm{do}_i(\alpha)] \mathbf{A}_i \mathtt{while}\, \varphi\, \mathtt{do}\, \alpha\, \mathtt{od}))$
6. $\models^{\mathbf{0}} \mathbf{A}_i \mathtt{while}\, \varphi\, \mathtt{do}\, \alpha\, \mathtt{od} \leftrightarrow (\neg\varphi \vee (\varphi \wedge \mathbf{A}_i \alpha \wedge \langle \mathrm{do}_i(\alpha) \rangle \mathbf{A}_i \mathtt{while}\, \varphi\, \mathtt{do}\, \alpha\, \mathtt{od}))$

Proof: All items are relatively straightforwardly proved. By way of illustration, we show the second and the last item. So let $i \in \mathrm{A}, \alpha, \alpha_1, \alpha_2 \in \mathrm{Ac}$ and $\varphi \in \mathrm{L}$ be arbitrary, and let $\mathrm{M} \in \mathbf{M}$ with state s be some model for L.

$$
\begin{aligned}
& \mathrm{M}, s \models^{\mathbf{1}} \mathbf{A}_i \alpha_1; \alpha_2 \\
\Leftrightarrow\ & \mathtt{c}^{\mathbf{1}}(i, \alpha_1; \alpha_2)(s) = \mathbf{1} \\
\Leftrightarrow\ & \mathtt{c}^{\mathbf{1}}(i, \alpha_1)(s) = \mathbf{1} \,\&\, \mathtt{c}^{\mathbf{1}}(i, \alpha_2)(\mathtt{r}^{\mathbf{1}}(i, \alpha_1)(s)) = \mathbf{1}
\end{aligned}
$$

$\Leftrightarrow \mathtt{c}^{\mathbf{1}}(i,\alpha_1)(s) = \mathbf{1} \,\&\, \forall s' \in \mathrm{S}(s' = \mathtt{r}^{\mathbf{1}}(i,\alpha_1)(s) \Rightarrow \mathtt{c}^{\mathbf{1}}(i,\alpha_2)(s') = \mathbf{1})$
$\Leftrightarrow \mathrm{M}, s \models^{\mathbf{1}} \mathbf{A}_i\alpha_1 \,\&\, \forall s' \in \mathrm{S}(s' = \mathtt{r}^{\mathbf{1}}(i,\alpha_1)(s) \Rightarrow \mathrm{M}, s' \models^{\mathbf{1}} \mathbf{A}_i\alpha_2)$
$\Leftrightarrow \mathrm{M}, s \models^{\mathbf{1}} \mathbf{A}_i\alpha_1 \,\&\, \mathrm{M}, s \models^{\mathbf{1}} [\mathrm{do}_i(\alpha_1)]\mathbf{A}_i\alpha_2$
$\Leftrightarrow \mathrm{M}, s \models^{\mathbf{1}} \mathbf{A}_i\alpha_1 \wedge [\mathrm{do}_i(\alpha_1)]\mathbf{A}_i\alpha_2$

$\mathrm{M}, s \models^{\mathbf{0}} \mathbf{A}_i\mathtt{while}\,\varphi\,\mathtt{do}\,\alpha\,\mathtt{od}$
$\Leftrightarrow \mathtt{c}^{\mathbf{0}}(i, \mathtt{while}\,\varphi\,\mathtt{do}\,\alpha\,\mathtt{od})(s) = \mathbf{1}$
$\Leftrightarrow \exists k \in \mathbb{N}(\mathtt{c}^{\mathbf{0}}(i, (\mathtt{confirm}\,\varphi;\alpha)^k; \mathtt{confirm}\,\neg\varphi)(s) = \mathbf{1})$
$\Leftrightarrow \mathtt{c}^{\mathbf{0}}(i, (\mathtt{confirm}\,\varphi;\alpha)^0; \mathtt{confirm}\,\neg\varphi)(s) = \mathbf{1}$ or
$\exists k \in \mathbb{N}(\mathtt{c}^{\mathbf{0}}(i, (\mathtt{confirm}\,\varphi;\alpha); (\mathtt{confirm}\,\varphi;\alpha)^k; \mathtt{confirm}\,\neg\varphi)(s) = \mathbf{1})$
$\Leftrightarrow \mathtt{c}^{\mathbf{0}}(i, \mathtt{skip}; \mathtt{confirm}\,\neg\varphi)(s) = \mathbf{1}$ or
$(\exists k \in \mathbb{N}(\mathtt{c}^{\mathbf{0}}(i, \mathtt{confirm}\,\varphi;\alpha)(s) = \mathbf{1} \,\&$
$\mathtt{c}^{\mathbf{0}}(i, (\mathtt{confirm}\,\varphi;\alpha)^k; \mathtt{confirm}\,\neg\varphi)(\mathtt{r}^{\mathbf{0}}(i, \mathtt{confirm}\,\varphi;\alpha)(s)) = \mathbf{1}))$
$\Leftrightarrow \mathtt{c}^{\mathbf{0}}(i, \mathtt{skip}) = \mathbf{1} \,\&\, \mathtt{c}^{\mathbf{0}}(i, \mathtt{confirm}\,\neg\varphi)(\mathtt{r}^{\mathbf{0}}(i, \mathtt{skip})(s)) = \mathbf{1}$ or
$(\mathtt{c}^{\mathbf{0}}(i, \mathtt{confirm}\,\varphi)(s) = \mathbf{1} \,\&\, \mathtt{c}^{\mathbf{0}}(i,\alpha)(\mathtt{r}^{\mathbf{0}}(i, \mathtt{confirm}\,\varphi)(s)) = \mathbf{1} \,\&$
$\mathtt{c}^{\mathbf{0}}(i, \mathtt{while}\,\varphi\,\mathtt{do}\,\alpha\,\mathtt{od})(\mathtt{r}^{\mathbf{0}}(i, \mathtt{confirm}\,\varphi;\alpha)(s)) = \mathbf{1})$
$\Leftrightarrow \mathtt{c}^{\mathbf{0}}(i, \mathtt{confirm}\,\neg\varphi)(s) = \mathbf{1}$ or
$(\mathrm{M}, s \models^{\mathbf{0}} \varphi \,\&\, \mathtt{c}^{\mathbf{0}}(i,\alpha)(s) = \mathbf{1} \,\&\, \mathtt{c}^{\mathbf{0}}(i, \mathtt{while}\,\varphi\,\mathtt{do}\,\alpha\,\mathtt{od})(\mathtt{r}^{\mathbf{0}}(i,\alpha)(s)) = \mathbf{1})$
$\Leftrightarrow \mathrm{M}, s \models^{\mathbf{0}} \neg\varphi$ or $(\mathrm{M}, s \models^{\mathbf{0}} \varphi \wedge \mathbf{A}_i\alpha \,\&\, \exists s' \in \mathrm{S}(s' = \mathtt{r}^{\mathbf{0}}(i,\alpha)(s) \,\&$
$\mathtt{c}^{\mathbf{0}}(i, \mathtt{while}\,\varphi\,\mathtt{do}\,\alpha\,\mathtt{od})(s') = \mathbf{1}))$
$\Leftrightarrow \mathrm{M}, s \models^{\mathbf{0}} \neg\varphi$ or $(\mathrm{M}, s \models^{\mathbf{0}} \varphi \wedge \mathbf{A}_i\alpha \,\&\, \mathrm{M}, s \models^{\mathbf{0}} \langle\mathrm{do}_i(\alpha)\rangle\mathbf{A}_i\mathtt{while}\,\varphi\,\mathtt{do}\,\alpha\,\mathtt{od})$
$\Leftrightarrow \mathrm{M}, s \models^{\mathbf{0}} \neg\varphi \vee (\varphi \wedge \mathbf{A}_i\alpha \wedge \langle\mathrm{do}_i(\alpha)\rangle\mathbf{A}_i\mathtt{while}\,\varphi\,\mathtt{do}\,\alpha\,\mathtt{od})$

⊠

3.10. Proposition. *For atomic actions $a \in \mathrm{At}$, the following correspondences hold in the class $\mathbf{F}$ of frames for $\mathbf{M}$ both for $\mathbf{b} = \mathbf{1}$ and $\mathbf{b} = \mathbf{0}$. The left-hand side of these correspondences is to be understood as a schema in $i \in \mathrm{A}, \varphi \in \mathrm{L}$.*

1. $\mathbf{K}_i[\mathrm{do}_i(a)]\varphi \to [\mathrm{do}_i(a)]\mathbf{K}_i\varphi \sim^{\mathbf{b}}$
$\forall s_0 \in \mathrm{S} \forall s_1 \in \mathrm{S}(\exists s_2 \in \mathrm{S}(s_2 = \mathtt{r}_0(i,a)(s_0) \,\&\, (s_2, s_1) \in \mathrm{R}(i)) \Rightarrow$
$\exists s_3 \in \mathrm{S}((s_0, s_3) \in \mathrm{R}(i) \,\&\, s_1 = \mathtt{r}_0(i,a)(s_3)))$
2. $\langle\mathrm{do}_i(a)\rangle\varphi \to [\mathrm{do}_i(a)]\varphi \sim^{\mathbf{b}}$
$\forall s \in \mathrm{S} \forall s' \in \mathrm{S} \forall s'' \in \mathrm{S}(\mathtt{r}_0(i,a)(s) = s' \,\&\, \mathtt{r}_0(i,a)(s) = s'' \Rightarrow s' = s'')$
3. $[\mathrm{do}_i(a;a)]\varphi \leftrightarrow [\mathrm{do}_i(a)]\varphi \sim^{\mathbf{b}} \forall s \in \mathrm{S}(\mathtt{r}_0(i,a)(\mathtt{r}_0(i,a)(s)) = \mathtt{r}_0(i,a)(s))$
4. $\langle\mathrm{do}_i(a)\rangle\top \sim^{\mathbf{b}} \forall s \in \mathrm{S}(\mathtt{r}_0(i,a)(s) \neq \emptyset)$
5. $\mathbf{A}_i a \to \langle\mathrm{do}_i(a)\rangle\top \sim^{\mathbf{b}} \forall s \in \mathrm{S}(\mathtt{c}_0(i,a)(s) = \mathbf{1} \Rightarrow \mathtt{r}_0(i,a)(s) \neq \emptyset)$

Proof: We show the first and the last item; the other items are less interesting and more easily proved. So let $\mathrm{F} = \langle \mathrm{S}, \mathrm{R}, \mathtt{r}_0, \mathtt{c}_0 \rangle$ be an arbitrary frame from $\mathbf{F}$. We have to show for the correspondences given above that $\mathrm{F} \models^{\mathbf{b}} \varphi$, where φ is the schema given

at the left-hand side, iff F satisfies the first-order property given at the right-hand side. Both items are shown by proving two implications, the latter of which is shown in its contrapositive form.

'$\Leftarrow$' Suppose that $\forall s_0 \in \mathrm{S} \forall s_1 \in \mathrm{S}(\exists s_2 \in \mathrm{S}(s_2 = \mathtt{r}_0(i,a)(s_0) \,\&\, (s_2, s_1) \in \mathrm{R}(i)) \Rightarrow \exists s_3 \in \mathrm{S}((s_0, s_3) \in \mathrm{R}(i) \,\&\, s_1 = \mathtt{r}_0(i,a)(s_3)))$. Now let $\pi : \Pi \times \mathrm{S} \to$ bool and $s_0 \in \mathrm{S}$ be arbitrary, and assume that $(\mathrm{F}, \pi), s_0 \models^{\mathbf{b}} \mathbf{K}_i[\mathrm{do}_i(a)]\varphi$, for some $i \in \mathrm{A}$ and $\varphi \in \mathrm{L}$. Suppose that $s_2 = \mathtt{r}_0(i,a)(s_0)$ and $(s_2, s_1) \in \mathrm{R}(i)$, for some $s_1, s_2 \in \mathrm{S}$. Then some $s_3 \in \mathrm{S}$ exists such that $(s_0, s_3) \in \mathrm{R}(i)$ and $\mathtt{r}_0(i,a)(s_3) = s_1$. Since $(\mathrm{F}, \pi), s_0 \models^{\mathbf{b}} \mathbf{K}_i[\mathrm{do}_i(a)]\varphi$, it then follows that $(\mathrm{F}, \pi), s_1 \models^{\mathbf{b}} \varphi$. Since both s_1 and s_2 are arbitrary, $(\mathrm{F}, \pi), s \models^{\mathbf{b}} \varphi$ for all s such that $s' = \mathtt{r}_0(i,a)(s_0)$ and $(s', s) \in \mathrm{R}(i)$, for some $s' \in \mathrm{S}$. Hence $(\mathrm{F}, \pi), s' \models^{\mathbf{b}} \mathbf{K}_i\varphi$ for all $s' = \mathtt{r}_0(i,a)(s)$, and thus $(\mathrm{F}, \pi), s \models^{\mathbf{b}} [\mathrm{do}_i(a)]\mathbf{K}_i\varphi$, which was to be shown.

'$\Rightarrow$' Suppose that not for all $s_0, s_1 \in \mathrm{S}$ holds that $(\exists s_2 \in \mathrm{S}(s_2 = \mathtt{r}_0(i,a)(s_0) \,\&\, (s_2, s_1) \in \mathrm{R}(i)) \Rightarrow \exists s_3 \in \mathrm{S}((s_0, s_3) \in \mathrm{R}(i) \,\&\, s_1 = \mathtt{r}_0(i,a)(s_3)))$. That is, some s_0, s_1 and s_2 exist such that $s_2 = \mathtt{r}_0(i,a)(s_0)$, $(s_2, s_1) \in \mathrm{R}(i)$ and for all $s_3 \in \mathrm{S}$ either $(s_0, s_3) \notin \mathrm{R}(i)$ or $s_1 \neq \mathtt{r}_0(i,a)(s_3)$. Now let $\pi : \Pi \times \mathrm{S} \to$ bool be such that $\forall s \in \mathrm{S}(\pi(p, s) = \mathbf{0} \Leftrightarrow s = s_2)$. Then $(\mathrm{F}, \pi), s_0 \models^{\mathbf{b}} \mathbf{K}_i[\mathrm{do}_i(a)]p$ and $(\mathrm{F}, \pi), s_0 \not\models^{\mathbf{b}} [\mathrm{do}_i(a)]\mathbf{K}_i p$, and thus $\mathrm{F} \not\models^{\mathbf{b}} \mathbf{K}_i[\mathrm{do}_i(a)]p \to [\mathrm{do}_i(a)]\mathbf{K}_i p$.

From cases '$\Leftarrow$' and '$\Rightarrow$' as given above we conclude that the first item of Proposition 3.10 indeed holds.

'$\Leftarrow$' Suppose that $\forall s \in \mathrm{S}(\mathtt{c}_0(i,a)(s) = \mathbf{1} \Rightarrow \mathtt{r}_0(i,a)(s) \neq \emptyset)$. Let $\pi : \Pi \times \mathrm{S} \to$ bool and $s \in \mathrm{S}$ be arbitrary, and assume that $(\mathrm{F}, \pi), s \models^{\mathbf{b}} \mathbf{A}_i a$. Then by definition, $\mathtt{c}_0(i,a)(s) = \mathbf{1}$, and hence $\mathtt{r}_0(i,a)(s) \neq \emptyset$. Thus $(\mathrm{F}, \pi), s \models^{\mathbf{b}} \langle \mathrm{do}_i(a) \rangle \top$, which was to be shown.

'$\Rightarrow$' Suppose not for all $s \in \mathrm{S}$ holds that $\mathtt{c}_0(i,a)(s) = \mathbf{1} \Rightarrow \mathtt{r}_0(i,a)(s) \neq \emptyset$, i.e. some $s \in \mathrm{S}$ exists such that $\mathtt{c}_0(i,a)(s) = \mathbf{1}$ and $\mathtt{r}_0(i,a)(s) = \emptyset$. For this s it holds that $(\mathrm{F}, \pi), s \models^{\mathbf{b}} \mathbf{A}_i a \wedge \neg\langle \mathrm{do}_i(a) \rangle \top$, for all valuations π, and thus $\mathrm{F} \not\models^{\mathbf{b}} \mathbf{A}_i a \to \langle \mathrm{do}_i(a) \rangle \top$.

It follows that the last item of Proposition 3.10 also holds.

⊠

3.11. Proposition. *The following lifting results hold for all* $\mathrm{F} \in \mathbf{F}$ *and* $\mathbf{b} \in$ bool*:*

- $\forall a \in \mathrm{At}(a$ *is A-realisable for* $\models^{\mathbf{b}}$ *in* $\mathrm{F}) \Rightarrow \forall \alpha \in \mathrm{Ac}(\alpha$ *is A-realisable for* $\models^{\mathbf{b}}$ *in* $\mathrm{F})$
- $\forall a \in \mathrm{At}(a$ *is deterministic for* $\models^{\mathbf{b}}$ *in* $\mathrm{F}) \Rightarrow \forall \alpha \in \mathrm{Ac}(\alpha$ *is deterministic for* $\models^{\mathbf{b}}$ *in* $\mathrm{F})$

Proof: Both cases are relatively easy shown by induction on the structure of actions.

⊠

3.6.1 A proof of soundness and completeness

Below we prove the soundness and completeness of deducibility in $\mathrm{LCap}_{\mathbf{b}}$ for $\models^{\mathbf{b}}$-validity in $\mathbf{M}$. As far as we know, this is one of the very few proofs of completeness that concerns a proof system in which both knowledge and actions are dealt with, and it is probably the very first in which abilities are also taken into consideration.

Rather than restricting ourselves to $\mathrm{LCap}_{\mathbf{b}}$ we will for the greater part consider general logics, cumulating in a very general and rather powerful result from which the soundness and completeness proof for $\mathrm{LCap}_{\mathbf{b}}$ can be derived as a corollary. Globally, the proof given below can be split into three parts. In the first part of the proof, canonical models are constructed for the logics induced by the proof systems $\Sigma_{\mathbf{1}}$ and $\Sigma_{\mathbf{0}}$. The possible worlds of these canonical models are given by so-called maximal theories. In the second part, the truth-theorem is proved, which states that truth in a possible world of a canonical model corresponds to being an element of the maximal theory that constitutes the possible world. In the last, and almost trivial, part of the proof it is shown how the general truth-theorem implies soundness and completeness of $\mathrm{LCap}_{\mathbf{b}}$ for $\mathbf{b}$-validity in $\mathbf{M}$.

The definition of canonical models as we give it is, as far as actions and dynamic constructs are concerned, based on the construction given by Goldblatt [41]. The proof of the truth-theorem is inspired by the one given by Spruit [123] to show completeness of the Segerberg axiomatisation for propositional dynamic logic. Due to the fact that formulae and actions are strongly related, the subformula or subaction relation is not adequate to apply induction upon in the proof of the truth-theorem. Instead a fairly complex ordering is used, well-foundedness of which is proved using some very powerful (and partly automated) techniques that are well-known from the theory of Term Rewriting Systems [24, 69].

Some preliminary definitions, propositions and lemmas are needed before the canonical models can be constructed.

3.32. Proposition. *For all* $\mathrm{M} \in \mathbf{M}$ *with state* s *and all* $i \in \mathrm{A}$, $\alpha \in \mathrm{Ac}$, $\varphi \in \mathrm{L}$ *and* $\phi \in \mathrm{Afm(L)}$ *we have:*

- $\mathrm{M}, s \models^{\mathbf{b}} \phi([\mathrm{do}_i(\mathtt{while}\,\varphi\,\mathtt{do}\,\alpha\,\mathtt{od})]\psi)$ *iff for all* $l \in \mathbb{N}$, $\mathrm{M}, s \models^{\mathbf{b}} \phi(\psi_l(i, \alpha, \varphi))$
- $\mathrm{M}, s \models^{\mathbf{b}} \phi(\neg \mathbf{A}_i \mathtt{while}\,\varphi\,\mathtt{do}\,\alpha\,\mathtt{od})$ *iff for all* $l \in \mathbb{N}$, $\mathrm{M}, s \models^{\mathbf{b}} \phi(\neg(\varphi_l(i, \alpha)))$

Proof: We prove both items by induction on the structure of ϕ.

- Let $\mathrm{M} \in \mathbf{M}$ with state s, and $i \in \mathrm{A}$, $\varphi, \psi \in \mathrm{L}$ and $\alpha \in \mathrm{Ac}$ be arbitrary.
  1. $\phi = \#$:

$$\begin{aligned} & \mathrm{M}, s \models^{\mathbf{b}} [\mathrm{do}_i(\mathtt{while}\,\varphi\,\mathtt{do}\,\alpha\,\mathtt{od})]\psi \\ \Leftrightarrow\ & \mathrm{M}, t \models^{\mathbf{b}} \psi \text{ for all } t \in \mathrm{S} \text{ such that } t = \mathtt{r}^{\mathbf{b}}(i, \mathtt{while}\,\varphi\,\mathtt{do}\,\alpha\,\mathtt{od})(s) \\ \Leftrightarrow\ & \mathrm{M}, t \models^{\mathbf{b}} \psi \text{ for all } t \in \mathrm{S} \text{ such that } t = \mathtt{r}^{\mathbf{b}}(i, (\mathtt{confirm}\,\varphi; \alpha)^l; \mathtt{confirm}\,\neg\varphi)(s) \end{aligned}$$

for all $l \in \mathbb{N}$
$\Leftrightarrow \mathrm{M}, s \models^{\mathbf{b}} [\mathrm{do}_i(\mathtt{confirm}\,\varphi; \alpha)^l; \mathtt{confirm}\,\neg\varphi)]\psi$ for all $l \in \mathbb{N}$
$\Leftrightarrow \mathrm{M}, s \models^{\mathbf{b}} \psi_l(i, \varphi, \alpha)$ for all $l \in \mathbb{N}$

2. $\phi = \mathbf{K}_i\phi'$:
$\mathrm{M}, s \models^{\mathbf{b}} (\mathbf{K}_i\phi')([\mathrm{do}_i(\mathtt{while}\,\varphi\,\mathtt{do}\,\alpha\,\mathtt{od})]\psi)$
$\Leftrightarrow \mathrm{M}, s \models^{\mathbf{b}} \mathbf{K}_i(\phi'([\mathrm{do}_i(\mathtt{while}\,\varphi\,\mathtt{do}\,\alpha\,\mathtt{od})]\psi))$
$\Leftrightarrow \mathrm{M}, t \models^{\mathbf{b}} \phi'([\mathrm{do}_i(\mathtt{while}\,\varphi\,\mathtt{do}\,\alpha\,\mathtt{od})]\psi)$ for all $t \in \mathrm{S}$ such that $(s,t) \in \mathrm{R}(i)$
$\Leftrightarrow \mathrm{M}, t \models^{\mathbf{b}} \phi'(\psi_l(i, \alpha, \varphi))$ for all $l \in \mathbb{N}$,
for all $t \in \mathrm{S}$ such that $(s,t) \in \mathrm{R}(i)$ (by induction hypothesis)
$\Leftrightarrow \mathrm{M}, s \models^{\mathbf{b}} \mathbf{K}_i\phi'(\psi_l(i, \alpha, \varphi))$ for all $l \in \mathbb{N}$
$\Leftrightarrow \mathrm{M}, s \models^{\mathbf{b}} (\mathbf{K}_i\phi')(\psi_l(i, \alpha, \varphi))$ for all $l \in \mathbb{N}$

3. The cases where $\phi = [\mathrm{do}_i(\beta)]\phi'$ and $\phi = \psi' \to \phi'$ are analogous to the case where $\phi = \mathbf{K}_i\phi'$.

- Let again $\mathrm{M} \in \mathbf{M}$ with state s, $i \in \mathrm{A}$, $\varphi \in \mathrm{L}$ and $\alpha \in \mathrm{Ac}$ be arbitrary.

1. $\phi = \#$:
$\mathrm{M}, s \models^{\mathbf{b}} \neg\mathbf{A}_i\mathtt{while}\,\varphi\,\mathtt{do}\,\alpha\,\mathtt{od}$
$\Leftrightarrow \mathsf{not}(\mathrm{M}, s \models^{\mathbf{b}} \mathbf{A}_i\mathtt{while}\,\varphi\,\mathtt{do}\,\alpha\,\mathtt{od})$
$\Leftrightarrow \mathsf{not}(\exists k \in \mathbb{N}(\mathtt{c}^{\mathbf{b}}(i, (\mathtt{confirm}\,\varphi; \alpha)^k; \mathtt{confirm}\,\neg\varphi)(s) = \mathbf{1}))$
$\Leftrightarrow \forall k \in \mathbb{N}(\mathsf{not}(\mathtt{c}^{\mathbf{b}}(i, (\mathtt{confirm}\,\varphi; \alpha)^k; \mathtt{confirm}\,\neg\varphi)(s) = \mathbf{1}))$
$\Leftrightarrow \forall k \in \mathbb{N}(\mathsf{not}(\mathrm{M}, s \models^{\mathbf{b}} \varphi_k(i, \alpha)))$
$\Leftrightarrow \forall k \in \mathbb{N}(\mathrm{M}, s \models^{\mathbf{b}} \neg(\varphi_k(i, \alpha_1)))$

2. $\phi = [\mathrm{do}_i(\beta)]\phi'$:
$\mathrm{M}, s \models^{\mathbf{b}} ([\mathrm{do}_i(\beta)]\phi')(\neg\mathbf{A}_i\mathtt{while}\,\varphi\,\mathtt{do}\,\alpha\,\mathtt{od})$
$\Leftrightarrow \mathrm{M}, s \models^{\mathbf{b}} [\mathrm{do}_i(\beta)](\phi'(\neg\mathbf{A}_i\mathtt{while}\,\varphi\,\mathtt{do}\,\alpha\,\mathtt{od}))$
$\Leftrightarrow \mathrm{M}, t \models^{\mathbf{b}} \phi'(\neg\mathbf{A}_i\mathtt{while}\,\varphi\,\mathtt{do}\,\alpha\,\mathtt{od})$ for all $t \in \mathrm{S}$ such that $t = \mathtt{r}^{\mathbf{b}}(i, \beta)(s)$
$\Leftrightarrow \mathrm{M}, t \models^{\mathbf{b}} \phi'(\neg(\varphi_l(i, \alpha)))$ for all $l \in \mathbb{N}$, for all $t \in \mathrm{S}$ such that $t = \mathtt{r}^{\mathbf{b}}(i, \beta)(s)$
(by induction hypothesis)
$\Leftrightarrow \mathrm{M}, s \models^{\mathbf{b}} [\mathrm{do}_i(\beta)](\phi'(\neg(\varphi_l(i, \alpha))))$ for all $l \in \mathbb{N}$
$\Leftrightarrow \mathrm{M}, s \models^{\mathbf{b}} ([\mathrm{do}_i(\beta)]\phi')(\neg(\varphi_l(i, \alpha)))$ for all $l \in \mathbb{N}$

3. The cases where $\phi = \mathbf{K}_i\phi'$ and $\phi = (\psi' \to \phi')$ are analogous to the case where $\phi = [\mathrm{do}_i(\beta)]\phi'$.

⊠

3.33. Proposition. *If* $\mathrm{M} \in \mathbf{M}$ *is a well-defined model from* $\mathbf{M}$*, then* $\Lambda^{\mathrm{M}}_{\mathbf{b}} \triangleq \{\varphi \in \mathrm{L} \mid \mathrm{M} \models^{\mathbf{b}} \varphi\}$ *is a* $\mathbf{b}$*-logic.*

Proof: We need to check for a given model $\mathrm{M} \in \mathbf{M}$ that the axioms of $\Sigma_{\mathbf{b}}$ are valid in M and that M is validity-preserving for the proof rules of $\Sigma_{\mathbf{b}}$. The validity of the axioms

A1–A9 and A12–A14 is easily checked. Axiom A10 follows from the determinism of all actions as stated in Corollary 3.12. Axiom $\mathrm{A15}_1$ is shown in Proposition 3.6, and $\mathrm{A15}_0$ is shown analogously. The validity-preservingness of M for the rules R1 and R2 follows from Proposition 3.32; M is easily seen to be validity-preserving for the other rules. As an example we show here the validity of axiom A10.

$$
\begin{aligned}
& \mathrm{M}, s \models^{\mathbf{b}} [\mathrm{do}_i(\mathtt{while}\,\varphi\,\mathtt{do}\,\alpha\,\mathtt{od})]\psi \\
\Leftrightarrow\ & \mathrm{M}, t \models^{\mathbf{b}} \psi \text{ for all } t \in \mathrm{S} \text{ such that } t = \mathtt{r}^{\mathbf{b}}(i, \mathtt{while}\,\varphi\,\mathtt{do}\,\alpha\,\mathtt{od})(s) \\
\Leftrightarrow\ & \mathrm{M}, t \models^{\mathbf{b}} \psi \text{ for all } t \in \mathrm{S} \text{ such that } t = \mathtt{r}^{\mathbf{b}}(i, \mathtt{confirm}\,\neg\varphi)(s) \text{ and} \\
& \mathrm{M}, t \models^{\mathbf{b}} \psi \text{ for all } t \in \mathrm{S} \text{ such that } t = \mathtt{r}^{\mathbf{b}}(i, (\mathtt{confirm}\,\varphi; \alpha); \mathtt{while}\,\varphi\,\mathtt{do}\,\alpha\,\mathtt{od})(s) \\
\Leftrightarrow\ & \mathrm{M}, s \models^{\mathbf{b}} [\mathrm{do}_i(\mathtt{confirm}\,\neg\varphi)]\psi \text{ and} \\
& \mathrm{M}, s \models^{\mathbf{b}} [\mathrm{do}_i((\mathtt{confirm}\,\varphi; \alpha); \mathtt{while}\,\varphi\,\mathtt{do}\,\alpha\,\mathtt{od})]\psi \\
\Leftrightarrow\ & \mathrm{M}, s \models^{\mathbf{b}} [\mathrm{do}_i(\mathtt{confirm}\,\neg\varphi)]\psi \wedge \\
& \qquad [\mathrm{do}_i(\mathtt{confirm}\,\varphi; \alpha)][\mathrm{do}_i(\mathtt{while}\,\varphi\,\mathtt{do}\,\alpha\,\mathtt{od})]\psi
\end{aligned}
$$

⊠

3.34. Proposition. *Let Λ be a logic. The following properties are shared by all Λ-theories Γ, for all $\varphi, \psi \in \mathrm{L}$, $i \in \mathrm{A}$, $\alpha \in \mathrm{Ac}$ and all $\phi \in \mathrm{Afm}(\mathrm{L})$:*

1. $\top \in \Gamma$
2. *if* $\Gamma \vdash^{\Lambda} \varphi$ *then* $\varphi \in \Gamma$
3. *if* $\vdash^{\Lambda} (\varphi \to \psi)$ *and* $\varphi \in \Gamma$ *then* $\psi \in \Gamma$
4. Γ *is* Λ*-consistent iff* $\bot \notin \Gamma$ *iff* $\Gamma \neq \mathrm{L}$
5. $(\varphi \wedge \psi) \in \Gamma$ *iff* $\varphi \in \Gamma$ *and* $\psi \in \Gamma$
6. *if* $\varphi \in \Gamma$ *or* $\psi \in \Gamma$ *then* $(\varphi \vee \psi) \in \Gamma$
7. $\phi([\mathrm{do}_i(\mathtt{while}\,\varphi\,\mathtt{do}\,\alpha\,\mathtt{od})]\psi) \in \Gamma$ *iff* $\{\phi(\psi_l(i, \varphi, \alpha)) \mid l \in \mathbb{N}\} \subseteq \Gamma$
8. $\phi(\neg \mathbf{A}_i \mathtt{while}\,\varphi\,\mathtt{do}\,\alpha\,\mathtt{od}) \in \Gamma$ *iff* $\{\phi(\neg(\varphi_l(i, \alpha))) \mid l \in \mathbb{N}\} \subseteq \Gamma$

Proof: The items 1 to 6 are fairly standard, and are proved by Goldblatt [41]. The cases 7 and 8 follow from the fact that theories contain the axioms A10 and $\mathrm{A15}_{\mathbf{b}}$ and are closed under $\Omega\mathrm{I}$ and $\Omega\mathrm{IA}$.
⊠

3.35. Definition. Let Λ be a logic. A maximal Λ-theory is a consistent Λ-theory Γ such that $\varphi \in \Gamma$ or $\neg\varphi \in \Gamma$ for all $\varphi \in \mathrm{L}$.

3.36. Proposition. *The following properties are shared by all maximal Λ-theories Γ, for Λ some logic, and $\varphi, \psi \in \mathrm{L}$.*

1. $\bot \notin \Gamma$
2. *exactly one of* φ *and* $\neg\varphi$ *belongs to* Γ*, for all* $\varphi \in \mathrm{L}$

3. $(\varphi \vee \psi) \in \Gamma$ *iff* $\varphi \in \Gamma$ *or* $\psi \in \Gamma$

3.37. Proposition. *For* Λ *a logic and all* $\varphi, \psi \in \mathrm{L}$, $\Phi, \Psi \subseteq \mathrm{L}$, $i \in \mathrm{A}$, $\alpha \in \mathrm{Ac}$ *and* $\phi \in \mathrm{Afm(L)}$ *we have:*

1. *if* $\varphi \in \Phi$ *then* $\Phi \vdash^{\Lambda} \varphi$
2. *if* $\Phi \vdash^{\Lambda} \varphi$ *and* $\Phi \subseteq \Psi$ *then* $\Psi \vdash^{\Lambda} \varphi$
3. $\vdash^{\Lambda} \varphi$ *iff* $\emptyset \vdash^{\Lambda} \varphi$
4. *if* $\Phi \vdash^{\Lambda} (\varphi \to \psi)$ *and* $\Phi \vdash^{\Lambda} \varphi$ *then* $\Phi \vdash^{\Lambda} \psi$
5. *if* $\Phi \vdash^{\Lambda} \phi(\psi_l(i, \varphi, \alpha))$ *for all* $l \in \mathbb{N}$ *then* $\Phi \vdash^{\Lambda} \phi([\mathrm{do}_i(\mathtt{while}\,\varphi\,\mathtt{do}\,\alpha\,\mathtt{od})]\psi)$
6. *if* $\Phi \vdash^{\Lambda} \phi(\neg(\varphi_l(i, \alpha)))$ *for all* $l \in \mathbb{N}$ *then* $\Phi \vdash^{\Lambda} \phi(\neg \mathbf{A}_i \mathtt{while}\,\varphi\,\mathtt{do}\,\alpha\,\mathtt{od})$

3.38. Theorem (The deduction theorem). *For* Λ *some logic and all* $\varphi, \psi \in \mathrm{L}$ *and* $\Phi \subseteq \mathrm{L}$ *we have that* $\Phi \cup \{\varphi\} \vdash^{\Lambda} \psi$ *iff* $\Phi \vdash^{\Lambda} (\varphi \to \psi)$.

Proof: We will prove the 'iff' by proving two implications:

'$\Leftarrow$' This case follows directly from items 1, 2, and 4 of Proposition 3.37.

'$\Rightarrow$' Assume that $\Phi \cup \{\varphi\} \vdash^{\Lambda} \psi$. Let $\Gamma \triangleq \{\rho \in \mathrm{L} \mid \Phi \vdash^{\Lambda} (\varphi \to \rho)\}$. We have to show that $\psi \in \Gamma$. For this it suffices to show that Γ is a Λ-theory containing $\Phi \cup \{\varphi\}$. We show here that Γ is closed under ΩIA; the proof of the other properties is easy and left to the reader. Assume that $\{\phi(\neg(\varphi'_l(i, \alpha))) \mid l \in \mathbb{N}\} \subseteq \Gamma$. Then $\Phi \vdash^{\Lambda} (\varphi \to \phi(\neg(\varphi'_l(i, \alpha))))$ for all $l \in \mathbb{N}$. Applying case 6 of Proposition 3.37 to the set $\{(\varphi \to \phi(\neg(\varphi'_l(i, \alpha)))) \mid l \in \mathbb{N}\}$ yields $\Phi \vdash^{\Lambda} (\varphi \to \phi(\neg \mathbf{A}_i \mathtt{while}\,\varphi'\,\mathtt{do}\,\alpha\,\mathtt{od}))$, hence $\phi(\neg \mathbf{A}_i \mathtt{while}\,\varphi'\,\mathtt{do}\,\alpha\,\mathtt{od}) \in \Gamma$. Thus Γ is closed under ΩIA.

$\boxtimes$

3.39. Corollary. *For* Λ *some logic and all* $\varphi \in \mathrm{L}$ *and* $\Phi \subseteq \mathrm{L}$ *we have:*

- $\Phi \cup \{\varphi\}$ *is* Λ*-consistent iff* $\Phi \nvdash^{\Lambda} \neg\varphi$
- $\Phi \cup \{\neg\varphi\}$ *is* Λ*-consistent iff* $\Phi \nvdash^{\Lambda} \varphi$

3.40. Definition. For $\Phi \subseteq \mathrm{L}$, $i \in \mathrm{A}$ and $\alpha \in \mathrm{Ac}$ we define:

- $\Phi/\mathbf{K}_i \triangleq \{\varphi \in \mathrm{L} \mid \mathbf{K}_i\varphi \in \Phi\}$
- $\mathbf{K}_i\Phi \triangleq \{\mathbf{K}_i\varphi \in \mathrm{L} \mid \varphi \in \Phi\}$
- $\Phi/[\mathrm{do}_i(\alpha)] \triangleq \{\varphi \in \mathrm{L} \mid [\mathrm{do}_i(\alpha)]\varphi \in \Phi\}$
- $[\mathrm{do}_i(\alpha)]\Phi \triangleq \{[\mathrm{do}_i(\alpha)]\varphi \in \mathrm{L} \mid \varphi \in \Phi\}$

3.41. Proposition. *For* Λ *some logic and all* $\varphi \in \mathrm{L}$, $i \in \mathrm{A}$ *and* $\Phi \subseteq \mathrm{L}$ *we have:*

- *if* $\Phi \vdash^{\Lambda} \varphi$ *then* $\mathbf{K}_i\Phi \vdash^{\Lambda} \mathbf{K}_i\varphi$
- *if* $\Phi \vdash^{\Lambda} \varphi$ *then* $[\mathrm{do}_i(\alpha)]\Phi \vdash^{\Lambda} [\mathrm{do}_i(\alpha)]\varphi$

PROOF: We show the first case; the second case is completely analogous. So let Γ be a Λ-theory such that $\mathbf{K}_i\Phi \subseteq \Gamma$. We need to show that $\mathbf{K}_i\varphi \in \Gamma$. Let $\Delta \triangleq \Gamma/\mathbf{K}_i$. Since $\Phi \vdash^\Lambda \varphi$, it suffices to show that Δ is a Λ-theory containing Φ. Then $\varphi \in \Delta$ and hence $\mathbf{K}_i\varphi \in \Gamma$.

1. $\Phi \subseteq \Delta$: If $\psi \in \Phi$, then $\mathbf{K}_i\psi \in \Gamma$ and hence $\psi \in \Delta$.
2. Δ contains Λ: If $\vdash^\Lambda \psi$, then by NK, $\vdash^\Lambda \mathbf{K}_i\psi$ and, since Γ is a Λ-theory, then $\mathbf{K}_i\psi \in \Gamma$, which implies $\psi \in \Delta$.
3. Δ is closed under MP, ΩI and ΩIA.
   - MP: If $\psi \in \Delta$ and $(\psi \to \psi_1) \in \Delta$, then $\mathbf{K}_i\psi \in \Gamma$ and $\mathbf{K}_i(\psi \to \psi_1) \in \Gamma$. Since Γ contains axiom A2, this implies $\mathbf{K}_i\psi_1 \in \Gamma$ and hence $\psi_1 \in \Delta$.
   - ΩI: If $\{\phi(\psi_l(j,\varphi',\alpha)) \mid l \in \mathbb{N}\} \subseteq \Delta$, then $\{\mathbf{K}_i\phi(\psi_l(j,\varphi',\alpha)) \mid l \in \mathbb{N}\} \subseteq \Gamma$. Applying ΩI to the set $\{\mathbf{K}_i\phi(\psi_l(j,\varphi',\alpha)) \mid l \in \mathbb{N}\}$ yields $\mathbf{K}_i\phi([\mathrm{do}_j(\texttt{while}\,\varphi'\,\texttt{do}\,\alpha\,\texttt{od})]\psi) \in \Gamma$, and hence $\phi([\mathrm{do}_j(\texttt{while}\,\varphi'\,\texttt{do}\,\alpha\,\texttt{od})]\psi) \in \Delta$.
   - ΩIA: If $\{\phi(\neg(\varphi'_l(j,\alpha))) \mid l \in \mathbb{N}\} \subseteq \Delta$, then $\{\mathbf{K}_i\phi(\neg(\varphi'_l(j,\alpha))) \mid l \in \mathbb{N}\} \subseteq \Gamma$. Applying ΩIA to $\{\mathbf{K}_i\phi(\neg(\varphi'_l(j,\alpha))) \mid l \in \mathbb{N}\}$ yields $\mathbf{K}_i\phi(\neg\mathbf{A}_j\texttt{while}\,\varphi'\,\texttt{do}\,\alpha\,\texttt{od}) \in \Gamma$, and hence $\phi(\neg\mathbf{A}_j\texttt{while}\,\varphi'\,\texttt{do}\,\alpha\,\texttt{od}) \in \Delta$.

   It follows that Δ is closed under MP, ΩI and ΩIA.

Since Δ contains Λ and is closed under MP, ΩI and ΩIA it follows that Δ is a Λ-theory. ⊠

3.42. COROLLARY. *Let Λ be some logic. For all Λ-theories Γ, and for $i \in \mathrm{A}$, $\alpha \in \mathrm{Ac}$ and $\varphi \in \mathrm{L}$ we have:*

- $\mathbf{K}_i\varphi \in \Gamma$ *iff* $\Gamma/\mathbf{K}_i \vdash^\Lambda \varphi$
- $[\mathrm{do}_i(\alpha)]\varphi \in \Gamma$ *iff* $\Gamma/[\mathrm{do}_i(\alpha)] \vdash^\Lambda \varphi$

3.43. PROPOSITION. *Let Λ be some logic. For all maximal Λ-theories Γ we have that if $\Gamma/[\mathrm{do}_i(\alpha)]$ is Λ-consistent then $\Gamma/[\mathrm{do}_i(\alpha)]$ is a maximal Λ-theory.*

PROOF: Suppose that $\Gamma/[\mathrm{do}_i(\alpha)]$ is Λ-consistent. We show that $\Gamma/[\mathrm{do}_i(\alpha)]$ is a Λ-theory and that for all $\varphi \in \mathrm{L}$, either $\varphi \in \Gamma/[\mathrm{do}_i(\alpha)]$ or $\neg\varphi \in \Gamma/[\mathrm{do}_i(\alpha)]$. Since by assumption $\Gamma/[\mathrm{do}_i(\alpha)]$ is consistent, this suffices to conclude that $\Gamma/[\mathrm{do}_i(\alpha)]$ is a maximal Λ-theory.

1. $\Gamma/[\mathrm{do}_i(\alpha)]$ contains Λ: If $\vdash^\Lambda \varphi$ then by NA, $\vdash^\Lambda [\mathrm{do}_i(\alpha)]\varphi$, and, since Γ is a Λ-theory, $[\mathrm{do}_i(\alpha)]\varphi \in \Gamma$. This implies that $\varphi \in \Gamma/[\mathrm{do}_i(\alpha)]$.
2. $\Gamma/[\mathrm{do}_i(\alpha)]$ is closed under MP, ΩI and ΩIA:
   - MP: Assume that $(\varphi \to \psi) \in \Gamma/[\mathrm{do}_i(\alpha)]$ and $\varphi \in \Gamma/[\mathrm{do}_i(\alpha)]$. Then $[\mathrm{do}_i(\alpha)](\varphi \to \psi) \in \Gamma$ and $[\mathrm{do}_i(\alpha)]\varphi \in \Gamma$, which implies, since Γ contains A6 and is closed under MP, that $[\mathrm{do}_i(\alpha)]\psi \in \Gamma$. This implies that $\psi \in \Gamma/[\mathrm{do}_i(\alpha)]$.
   - ΩI: If $\{\phi(\psi_l(j,\varphi,\beta)) \mid l \in \mathbb{N}\} \subseteq \Gamma/[\mathrm{do}_i(\alpha)]$, then $\{[\mathrm{do}_i(\alpha)]\phi(\psi_l(j,\varphi,\beta)) \mid l \in \mathbb{N}\} \subseteq \Gamma$. Applying ΩI to the set of afms $\{[\mathrm{do}_i(\alpha)]\phi(\psi_l(j,\varphi,\beta)) \mid l \in \mathbb{N}\}$, yields that

$[\mathrm{do}_i(\alpha)]\phi([\mathrm{do}_j(\texttt{while}\,\varphi\,\texttt{do}\,\beta\,\texttt{od})]\psi) \in \Gamma$, and hence $\phi([\mathrm{do}_j(\texttt{while}\,\varphi\,\texttt{do}\,\beta\,\texttt{od})]\psi) \in \Gamma/[\mathrm{do}_i(\alpha)]$.

- ΩIA: Let $\{\phi(\neg(\varphi_l(j,\beta))) \mid l \in \mathbb{N}\} \subseteq \Gamma/[\mathrm{do}_i(\alpha)]$. Then $\{[\mathrm{do}_i(\alpha)]\phi(\neg(\varphi_l(j,\beta))) \mid l \in \mathbb{N}\} \subseteq \Gamma$. Applying ΩI to the set $\{[\mathrm{do}_i(\alpha)]\phi(\neg(\varphi_l(j,\beta))) \mid l \in \mathbb{N}\}$ yields that $[\mathrm{do}_i(\alpha)]\phi(\neg\mathbf{A}_j\texttt{while}\,\varphi\,\texttt{do}\,\beta\,\texttt{od}) \in \Gamma$. Hence $\phi(\neg\mathbf{A}_j\texttt{while}\,\varphi\,\texttt{do}\,\beta\,\texttt{od}) \in \Gamma/[\mathrm{do}_i(\alpha)]$.

3. Since Γ is a theory, Γ contains axiom A11: $[\mathrm{do}_i(\alpha)]\varphi \vee [\mathrm{do}_i(\alpha)]\neg\varphi$ for all $i,\ \alpha$ and φ. Since Γ is maximal, $[\mathrm{do}_i(\alpha)]\varphi \in \Gamma$ or $[\mathrm{do}_i(\alpha)]\neg\varphi \in \Gamma$ for all $i,\ \alpha$ and φ. But this implies that $\varphi \in \Gamma/[\mathrm{do}_i(\alpha)]$ or $\neg\varphi \in \Gamma/[\mathrm{do}_i(\alpha)]$, for all $\varphi \in \mathrm{L}$.

By items 1, 2, and 3 it follows that $\Gamma/[\mathrm{do}_i(\alpha)]$ is a maximal Λ-theory if $\Gamma/[\mathrm{do}_i(\alpha)]$ is Λ-consistent.

⊠

3.44. Definition. For Λ some logic, the set S_Λ is defined by $\mathrm{S}_\Lambda \triangleq \{\Gamma \subseteq \mathrm{L} \mid \Gamma \text{ is a maximal } \Lambda\text{-theory}\}$.

3.45. Proposition. *For Λ some logic and all $\Phi \subseteq \mathrm{L}$ and $\varphi \in \mathrm{L}$ we have:*

- $\Phi \vdash^\Lambda \varphi$ *iff for all* $\Gamma \in \mathrm{S}_\Lambda$ *such that* $\Phi \subseteq \Gamma$ *holds that* $\varphi \in \Gamma$
- $\vdash^\Lambda \varphi$ *iff for all* $\Gamma \in \mathrm{S}_\Lambda$ *holds that* $\varphi \in \Gamma$

Proof: The second item follows by instantiating the first item with $\Phi = \emptyset$ and using item 3 of Proposition 3.37. We show the first item by proving two implications.

'$\Rightarrow$' By definition of $\Phi \vdash^\Lambda \varphi$.

'$\Leftarrow$' We show: if $\Phi \nvdash^\Lambda \varphi$ then some $\Gamma \in \mathrm{S}_\Lambda$ exists such that $\Phi \subseteq \Gamma$ and $\varphi \notin \Gamma$. We construct a Γ that satisfies this demand. To this end, we start by making an enumeration $\rho_0, \rho_1, \ldots$ of the formulae of L. Using this enumeration, the increasing sequence of sets $\Gamma_l \subseteq \mathrm{L}$ is for $l \in \mathbb{N}$ inductively defined as follows:

1. $\Gamma_0 = \Phi \cup \{\neg\varphi\}$
2. Assume that Γ_k has been defined. The set Γ_{k+1} is defined by the following algorithm, written in a high-level programming language pseudocode:

   if $\Gamma_k \vdash^\Lambda \rho_k$ then $\Gamma_{k+1} = \Gamma_k \cup \{\rho_k\}$
   elsif ρ_k is of the form $\phi([\mathrm{do}_i(\texttt{while}\,\varphi\,\texttt{do}\,\alpha\,\texttt{od})]\psi)$
   then $\Gamma_{k+1} = \Gamma_k \cup \{\neg\phi(\psi_j(i,\varphi,\alpha))\} \cup \{\neg\rho_k\}$,
   where j is the least number such that $\Gamma_k \nvdash^\Lambda \phi(\psi_j(i,\varphi,\alpha))$
   (this j exists since otherwise application of ΩI would yield $\Gamma_k \vdash^\Lambda \rho_k$)
   elsif ρ_k is of the form $\phi(\neg\mathbf{A}_i\texttt{while}\,\varphi\,\texttt{do}\,\alpha\,\texttt{od})$
   then $\Gamma_{k+1} = \Gamma_k \cup \{\neg\phi(\neg(\varphi_j(i,\alpha)))\} \cup \{\neg\rho_k\}$,
   where j is the least number such that $\Gamma_k \nvdash^\Lambda \phi(\neg(\varphi_j(i,\alpha)))$
   (this j exists since otherwise application of ΩIA would yield $\Gamma_k \vdash^\Lambda \rho_k$)
   else $\Gamma_{k+1} = \Gamma_k \cup \{\neg\rho_k\}$
   fi

Now Γ is defined by $\Gamma \triangleq \cup_{l\in\mathbb{N}}\Gamma_l$. We show that Γ is a maximal Λ-theory.

3.46. LEMMA. *The set Γ_l is Λ-consistent for all $l \in \mathbb{N}$.*

PROOF: We prove the lemma by induction on l. Since $\Phi \nvdash^\Lambda \varphi$, we have that $\Phi \cup \{\neg\varphi\} = \Gamma_0$ is Λ-consistent by Corollary 3.39. Now assume that Γ_k is consistent. Consider the four possibilities for the definition of Γ_{k+1}:

1. If $\Gamma_k \vdash^\Lambda \rho_k$, then, since Γ_k is assumed to be Λ-consistent, $\Gamma_k \nvdash^\Lambda \neg\rho_k$, and hence, by Corollary 3.39, $\Gamma_{k+1} = \Gamma_k \cup \{\rho_k\}$ is Λ-consistent.
2. If $\Gamma_k \cup \{\neg\phi(\psi_j(i,\varphi,\alpha))\} \cup \{\neg\rho_k\}$ were to be Λ-inconsistent, we would have $\Gamma_k \cup \{\neg\phi(\psi_j(i,\varphi,\alpha))\} \vdash^\Lambda \rho_k$, where $\rho_k = \phi([\mathrm{do}_i(\texttt{while}\,\varphi\,\texttt{do}\,\alpha\,\texttt{od})]\psi)$. Since we have $\vdash^\Lambda \rho_k \to \phi(\psi_l(i,\varphi,\alpha))$ for all $l \in \mathbb{N}$, we also have $\Gamma_k \cup \{\neg\phi(\psi_j(i,\varphi,\alpha))\} \vdash^\Lambda \phi(\psi_j(i,\varphi,\alpha))$, which implies that $\Gamma_k \cup \{\neg\phi(\psi_j(i,\varphi,\alpha))\}$ is Λ-inconsistent. But then, by Corollary 3.39, $\Gamma_k \vdash^\Lambda \phi(\psi_j(i,\varphi,\alpha))$ which contradicts the fact that $\Gamma_k \nvdash^\Lambda \phi(\psi_j(i,\varphi,\alpha))$. Hence Γ_{k+1} is Λ-consistent.
3. If $\Gamma_k \cup \{\neg(\phi(\neg\varphi_j(i,\alpha)))\} \cup \{\neg\rho_k\}$ were to be Λ-inconsistent, we would have $\Gamma_k \cup \{\neg\phi(\neg(\varphi_j(i,\alpha)))\} \vdash^\Lambda \rho_k$, where $\rho_k = \phi(\neg\mathbf{A}_i\texttt{while}\,\varphi\,\texttt{do}\,\alpha\,\texttt{od})$. Since we have $\vdash^\Lambda \rho_k \to \phi(\neg(\varphi_l(i,\alpha)))$ for all $l \in \mathbb{N}$, we also have $\Gamma_k \cup \{\neg\phi(\neg(\varphi_j(i,\alpha)))\} \vdash^\Lambda \phi(\neg(\varphi_j(i,\alpha)))$, which implies that $\Gamma_k \cup \{\neg\phi(\neg(\varphi_j(i,\alpha)))\}$ is Λ-inconsistent. But then $\Gamma_k \vdash^\Lambda \phi(\neg(\varphi_j(i,\alpha)))$ which contradicts the fact that $\Gamma_k \nvdash^\Lambda \phi(\neg(\varphi_j(i,\alpha)))$. Hence Γ_{k+1} is Λ-consistent.
4. If $\Gamma_k \nvdash^\Lambda \rho_k$ then $\Gamma_k \cup \{\neg\rho_k\}$ is Λ-consistent by Corollary 3.39.

⊠

3.47. LEMMA. *The set Γ as constructed above is maximal, i.e. for all $\varphi \in \mathrm{L}$, exactly one of φ and $\neg\varphi$ is an element of Γ.*

PROOF: Let $\psi \in \mathrm{L}$ be arbitrary, then $\psi = \rho_k$ for some $k \in \mathbb{N}$. By construction, now either $\rho_k \in \Gamma_{k+1}$ or $\neg\rho_k \in \Gamma_{k+1}$, hence either $\psi \in \Gamma$ or $\neg\psi \in \Gamma$. Suppose both ψ and $\neg\psi$ in Γ. Then for some $k \in \mathbb{N}$, $\{\psi, \neg\psi\} \subseteq \Gamma_k$, which would make Γ_k inconsistent. Since this contradicts the result of Lemma 3.46 given above, it follows that ψ and $\neg\psi$ are not both in Γ.

⊠

3.48. LEMMA. *The set Γ as constructed above is a Λ-theory.*

PROOF: We need to show that Γ contains Λ and is closed under MP, ΩI, and ΩIA. So let $\varphi, \psi \in \mathrm{L}$, $i \in \mathrm{A}$ and $\alpha \in \mathrm{Ac}$ be arbitrary.

1. Γ contains Λ: If $\vdash^\Lambda \varphi$, where $\varphi = \rho_k$ for some $k \in \mathbb{N}$, then $\Gamma_k \vdash^\Lambda \rho_k$ and hence $\varphi = \rho_k \in \Gamma_{k+1} \subseteq \Gamma$.
2. Closure under MP, ΩI, and ΩIA:
   - MP: Suppose that $\varphi,\ \varphi \to \psi \in \Gamma$. If $\psi \notin \Gamma$, then $\neg\psi \in \Gamma$, since Γ is maximal by Lemma 3.47. Hence $\{\varphi, \varphi \to \psi, \neg\psi\} \in \Gamma_k$ for some $k \in \mathbb{N}$, which would make Γ_k Λ-inconsistent. This leads to a contradiction with Lemma 3.46, hence

$\psi \in \Gamma$.

- ΩI: Suppose $\{\phi(\psi_l(i,\varphi,\alpha)) \mid l \in \mathbb{N}\} \subseteq \Gamma$. Let $\phi([\mathrm{do}_i(\texttt{while}\,\varphi\,\texttt{do}\,\alpha\,\texttt{od})]\psi) = \rho_k$, for some $k \in \mathbb{N}$. If $\rho_k \notin \Gamma$, then $\Gamma_k \nvdash^\Lambda \rho_k$, and, by case 2 of the construction of Γ_{k+1}, this implies that $\neg\phi(\psi_j(i,\varphi,\alpha)) \in \Gamma_{k+1}$, where $j \in \mathbb{N}$ is the least number such that $\Gamma_k \nvdash^\Lambda \phi(\psi_j(i,\varphi,\alpha))$. Hence $\neg\phi(\psi_j(i,\varphi,\alpha)) \in \Gamma$, and by Lemma 3.47, $\phi(\psi_j(i,\varphi,\alpha)) \notin \Gamma$, which contradicts the assumption that $\{\phi(\psi_l(i,\varphi,\alpha)) \mid l \in \mathbb{N}\} \subseteq \Gamma$. Hence $\phi([\mathrm{do}_i(\texttt{while}\,\varphi\,\texttt{do}\,\alpha\,\texttt{od})]\psi) \in \Gamma$.
- ΩIA: Suppose $\{\phi(\neg(\varphi_l(i,\alpha))) \mid l \in \mathbb{N}\} \subseteq \Gamma$. Let $\phi(\neg\mathbf{A}_i\texttt{while}\,\varphi\,\texttt{do}\,\alpha\,\texttt{od}) = \rho_k$, for some $k \in \mathbb{N}$. If $\rho_k \notin \Gamma$, then $\Gamma_k \nvdash^\Lambda \rho_k$, and by case 2 of the construction of Γ_{k+1}, this implies that $\neg(\phi(\neg\varphi_j(i,\alpha))) \in \Gamma_{k+1}$, for $j \in \mathbb{N}$ the least number such that $\Gamma_k \nvdash^\Lambda \phi(\neg(\varphi_j(i,\alpha)))$. Hence $\neg\phi(\neg(\varphi_j(i,\alpha))) \in \Gamma$, and $\phi(\neg(\varphi_j(i,\alpha))) \notin \Gamma$, which contradicts the assumption that $\{\phi(\neg(\varphi_l(i,\alpha))) \mid l \in \mathbb{N}\} \subseteq \Gamma$. Hence $\phi(\neg\mathbf{A}_i\texttt{while}\,\varphi\,\texttt{do}\,\alpha\,\texttt{od}) \in \Gamma$.

We conclude that Γ is closed under MP, ΩI and ΩIA.

Since Γ contains Λ and is closed under MP, ΩI and ΩIA, we conclude that Γ is a Λ-theory.

⊠

Now if Γ is Λ-inconsistent, then $\Gamma \vdash^\Lambda \bot$. Since, by Lemma 3.48, Γ is a Λ-theory, it follows by Proposition 3.34(2) that $\bot \in \Gamma$. Then $\bot \in \Gamma_k$ for some $k \in \mathbb{N}$, which contradicts the Λ-consistency of Γ_k which was shown in Lemma 3.46. Hence Γ is a Λ-theory (Lemma 3.48) which is maximal (Lemma 3.47) and Λ-consistent, thus Γ is a maximal Λ-theory. Note that by construction of Γ, $\Phi \subseteq \Gamma$ and $\neg\varphi \in \Gamma$, which suffices to conclude the right-to-left implication.

⊠

3.49. Definition. Let Λ be some logic. The canonical model M_Λ for Λ is defined by $\mathrm{M}_\Lambda \triangleq \langle \mathrm{S}_\Lambda, \pi_\Lambda, \mathrm{R}_\Lambda, \mathtt{r}_\Lambda, \mathtt{c}_\Lambda\rangle$ where

1. S_Λ is the set of maximal Λ-theories
2. $\pi_\Lambda(p,s) = \mathbf{1}$ iff $p \in s$, for $p \in \Pi$ and $s \in \mathrm{S}_\Lambda$
3. $(s,t) \in \mathrm{R}_\Lambda(i)$ iff $s/\mathbf{K}_i \subseteq t$, for $s,t \in \mathrm{S}_\Lambda$ and $i \in \mathrm{A}$
4. $t = \mathtt{r}_\Lambda(i,a)(s)$ iff $s/\mathrm{do}_i(a) \subseteq t$, for $i \in \mathrm{A}$, $a \in \mathrm{At}$ and $s,t \in \mathrm{S}_\Lambda$
5. $\mathtt{c}_\Lambda(i,a)(s) = \mathbf{1}$ if $\mathbf{A}_i a \in s$ and $\mathtt{c}_\Lambda(i,a)(s) = \mathbf{0}$ if $\mathbf{A}_i a \notin s$ for $i \in \mathrm{A}$, $a \in \mathrm{At}$ and $s \in \mathrm{S}_\Lambda$

3.50. Proposition. *Let Λ be some logic. The canonical model M_Λ for Λ as defined above is a well-defined model from $\mathbf{M}$.*

Proof: Let Λ be some logic. In order to show that M_Λ is a well-defined model from $\mathbf{M}$ we have to show that the demands determining well-definedness of models are met by M_Λ. It is easily seen that S_Λ, π_Λ, and $\mathtt{c}_\Lambda$ are well-defined, which leaves to show that R_Λ

and $\mathtt{r}_\Lambda$ are. To prove that $\mathrm{R}(i)$ is an equivalence relation, assume that $i \in \mathrm{A}$ and that $\{s,t,u\} \subseteq \mathrm{S}_\Lambda$. We show:

1. $(s,s) \in \mathrm{R}(i)$, i.e. $\mathrm{R}(i)$ is reflexive.
2. if $(s,t) \in \mathrm{R}(i)$ and $(s,u) \in \mathrm{R}(i)$ then $(t,u) \in \mathrm{R}(i)$, i.e. $\mathrm{R}(i)$ is Euclidean.

To show the reflexivity of $\mathrm{R}(i)$, note that $(s,t) \in \mathrm{R}(i)$ iff $s/\mathbf{K}_i \subseteq t$. Now since s contains axiom A3: $\mathbf{K}_i\varphi \to \varphi$, we have for $\varphi \in s/\mathbf{K}_i$ that $\mathbf{K}_i\varphi \in s$ and hence $\varphi \in s$ by MP. Thus $s/\mathbf{K}_i \subseteq s$, hence $(s,s) \in \mathrm{R}(i)$. To show that $\mathrm{R}(i)$ is Euclidean assume that $\varphi \in t/\mathbf{K}_i$, i.e., $\mathbf{K}_i\varphi \in t$. To prove: $\varphi \in u$. Suppose $\varphi \notin u$. Then since $(s,u) \in \mathrm{R}(i)$, $\varphi \notin s/\mathbf{K}_i$, i.e., $\mathbf{K}_i\varphi \notin s$. Since s is a maximal Λ-theory this implies that $\neg\mathbf{K}_i\varphi \in s$, and since s contains axiom A5: $\neg\mathbf{K}_i\varphi \to \mathbf{K}_i\neg\mathbf{K}_i\varphi$, also $\mathbf{K}_i\neg\mathbf{K}_i\varphi \in s$. Since $(s,t) \in \mathrm{R}(i)$, $s/\mathbf{K}_i \subseteq t$ and thus $\neg\mathbf{K}_i\varphi \in t$. But then $\mathbf{K}_i\varphi \in t$ and $\neg\mathbf{K}_i\varphi \in t$ which contradicts the consistency of t. Thus $\mathrm{R}(i)$ is Euclidean, and, combined with the reflexivity, this ensures that $\mathrm{R}(i)$ is an equivalence relation.

To show that $\mathtt{r}_\Lambda$ is well-defined, it needs to be shown that for all $i \in \mathrm{A}$, $a \in \mathrm{At}$ and $s \in \mathrm{S}_\Lambda$ it holds that $\mathtt{r}_\Lambda(i,a)(s) \in \mathrm{S}_\Lambda$ or $\mathtt{r}_\Lambda(i,a)(s) = \emptyset$. To this end it suffices to show for arbitrary $i \in \mathrm{A}$, $a \in \mathrm{At}$ and $s,t,u \in \mathrm{S}_\Lambda$ that if $t = \mathtt{r}_\Lambda(i,a)(s)$ and $u = \mathtt{r}_\Lambda(i,a)(s)$ then $t = u$. By definition it follows that $s/[\mathrm{do}_i(a)] \subseteq t$ and $s/[\mathrm{do}_i(a)] \subseteq u$ if both $t = \mathtt{r}_\Lambda(i,a)(s)$ and $u = \mathtt{r}_\Lambda(i,a)(s)$. Since both t and u are maximal Λ-theories, both t and u are Λ-consistent, and hence $s/[\mathrm{do}_i(a)]$ is Λ-consistent. But then, by Proposition 3.43, $s/[\mathrm{do}_i(a)]$ is a maximal Λ-theory, which is properly contained only in L. Hence $s/[\mathrm{do}_i(a)] = t$ and $s/[\mathrm{do}_i(a)] = u$, which suffices to conclude that $\mathtt{r}_\Lambda$ is well-defined.
⊠

Up till now, the two proof systems $\Sigma_\mathbf{0}$ and $\Sigma_\mathbf{1}$ were dealt with identically, i.e. in none of the definitions or propositions given above one needs to distinguish the proof systems or the logics based on these proof systems. From this point on, however, we need to treat the two systems, and thereby the logics, differently. We start with finishing the proof of soundness and completeness for **1**-logics, and indicate thereafter how this proof needs to be modified to end up with one for **0**-logics.

The presence of the confirmation action, which tightly links actions and formulae, makes that the subformula- or subaction-relation is not adequate to apply induction upon in the proof of the truth-theorem, the theorem which links satisfiability in a state of the canonical model to being an element of the maximal theory which constitutes the state. Instead we need a more elaborate relation, which is defined below.

3.51. Definition. The relation $\prec$ is the smallest relation on $\{0,1\} \times \mathrm{L}$ that satisfies for all $\varphi,\psi \in \mathrm{L}$, $i \in \mathrm{A}$, and $\alpha,\alpha_1,\alpha_2 \in \mathrm{Ac}$ the following constraints:

1. $(0,\varphi) \prec (0,\varphi \vee \psi)$
2. $(0,\psi) \prec (0,\varphi \vee \psi)$
3. $(0,\varphi) \prec (0,\neg\varphi)$

4. $(0, \varphi) \prec (0, \mathbf{K}_i\varphi)$
5. $(0, \varphi) \prec (0, [\mathrm{do}_i(\alpha)]\varphi)$
6. $(1, [\mathrm{do}_i(\alpha)]\varphi) \prec (0, [\mathrm{do}_i(\alpha)]\varphi)$
7. $(1, [\mathrm{do}_i(\alpha_1)][\mathrm{do}_i(\alpha_2)]\varphi) \prec (1, [\mathrm{do}_i(\alpha_1; \alpha_2)]\varphi)$
8. $(1, [\mathrm{do}_i(\alpha_2)]\varphi) \prec (1, [\mathrm{do}_i(\alpha_1; \alpha_2)]\varphi)$
9. $(1, [\mathrm{do}_i(\mathtt{confirm}\,\varphi; \alpha_1)]\psi) \prec (1, [\mathrm{do}_i(\mathtt{if}\,\varphi\,\mathtt{then}\,\alpha_1\,\mathtt{else}\,\alpha_2\,\mathtt{fi})]\psi)$
10. $(1, [\mathrm{do}_i(\mathtt{confirm}\,\neg\varphi; \alpha_2)]\psi) \prec (1, [\mathrm{do}_i(\mathtt{if}\,\varphi\,\mathtt{then}\,\alpha_1\,\mathtt{else}\,\alpha_2\,\mathtt{fi})]\psi)$
11. $(1, \psi_l(i, \varphi, \alpha)) \prec (1, [\mathrm{do}_i(\mathtt{while}\,\varphi\,\mathtt{do}\,\alpha\,\mathtt{od})]\psi)$ for all $l \in \mathbb{N}$
12. $(0, \neg\varphi) \prec (1, [\mathrm{do}_i(\mathtt{confirm}\,\varphi)]\psi)$
13. $(1, \mathbf{A}_i\alpha) \prec (0, \mathbf{A}_i\alpha)$
14. $(1, \mathbf{A}_i\alpha_1) \prec (1, \mathbf{A}_i\alpha_1; \alpha_2)$
15. $(1, \mathbf{A}_i\alpha_2) \prec (1, \mathbf{A}_i\alpha_1; \alpha_2)$
16. $(1, [\mathrm{do}_i(\alpha_1)]\mathbf{A}_i\alpha_2) \prec (1, \mathbf{A}_i\alpha_1; \alpha_2)$
17. $(1, \mathbf{A}_i\mathtt{confirm}\,\varphi; \alpha_1) \prec (1, \mathbf{A}_i\mathtt{if}\,\varphi\,\mathtt{then}\,\alpha_1\,\mathtt{else}\,\alpha_2\,\mathtt{fi})$
18. $(1, \mathbf{A}_i\mathtt{confirm}\,\neg\varphi; \alpha_2) \prec (1, \mathbf{A}_i\mathtt{if}\,\varphi\,\mathtt{then}\,\alpha_1\,\mathtt{else}\,\alpha_2\,\mathtt{fi})$
19. $(1, \varphi_l(i, \alpha)) \prec (1, \mathbf{A}_i\mathtt{while}\,\varphi\,\mathtt{do}\,\alpha\,\mathtt{od})$ for all $l \in \mathbb{N}$
20. $(0, \varphi) \prec (1, \mathbf{A}_i\mathtt{confirm}\,\varphi)$

3.52. Definition. The ordering $<$ is defined as the transitive closure of $\prec$, and $\leq$ is defined as the reflexive closure of $<$.

3.53. Proposition. *The ordering $<$ is well-founded.*

Proof: The proof of this proposition is quite elaborate; it can be found in [52] where it takes over three pages. Basically, the idea is to use a powerful technique well-known from the theory of Term Rewriting Systems, viz. the lexicographic path ordering. Using this technique it suffices to select an appropriate well-founded precedence on the function symbols of the language in order to conclude that the ordering $\prec$ is well-founded. Since the actual proof is not only rather elaborate but also contains many details that are completely outside the scope of this thesis, it is omitted here; those who are interested can find all details in [52].
⊠

Having proved that the ordering $<$ is well-founded, we can use it in the proof of the truth-theorem.

3.54. Theorem (The truth-theorem). *Let Λ be some* **1***-logic. For any $\varphi \in \mathrm{L}$, and any $s \in \mathrm{S}_\Lambda$ we have: $\mathrm{M}_\Lambda, s \models^{\mathbf{1}} \varphi$ iff $\varphi \in s$.*

Proof: We prove the theorem by proving the following (stronger) properties for all $\varphi, \psi \in \mathrm{L}$, $i \in \mathrm{A}$, $\alpha \in \mathrm{Ac}$, and $s \in \mathrm{S}_\Lambda$:

1. For all $(0,\psi) \leq (0,\varphi)$ we have: $\mathrm{M}_\Lambda, s \models^{\mathbf{1}} \psi$ iff $\psi \in s$
2. For all $(1,[\mathrm{do}_i(\alpha)]\psi) < (0,\varphi)$ we have:
   (a) $\psi \in t$ for $t = \mathtt{r}^{\mathbf{1}}(i,\alpha)(s) \Rightarrow [\mathrm{do}_i(\alpha)]\psi \in s$
   (b) if $t = \mathtt{r}^{\mathbf{1}}(i,\alpha)(s)$ and $[\mathrm{do}_i(\alpha)]\psi \in s$ then $\psi \in t$
3. For all $(1,\mathbf{A}_i\alpha) < (0,\varphi)$ we have: $\mathtt{c}^{\mathbf{1}}(i,\alpha)(s) = \mathbf{1}$ iff $\mathbf{A}_i\alpha \in s$

where $\mathtt{r}^{\mathbf{1}}$ and $\mathtt{c}^{\mathbf{1}}$ are the functions induced by $\mathtt{r}_\Lambda$ and $\mathtt{c}_\Lambda$ in the way described in Definition 3.3. The theorem then follows from the first item, since $(0,\varphi) \leq (0,\varphi)$. So let $\varphi \in \mathrm{L}$ be some fixed formula. We start by proving the first property. Let $\psi \in \mathrm{L}$ be such that $(0,\psi) \leq (0,\varphi)$. Consider the various cases for ψ:

- $\psi = p$, for $p \in \Pi$. By definition of π_Λ we have that $\pi_\Lambda(p,s) = \mathbf{1}$ iff $p \in s$.
- $\psi = \psi_1 \wedge \psi_2$. Since $(0,\psi_1) < (0,\psi_1 \wedge \psi_2)$ and $(0,\psi_2) < (0,\psi_1 \wedge \psi_2)$, we have that $\mathrm{M}_\Lambda, s \models^{\mathbf{1}} \psi_1 \wedge \psi_2$ iff ($\mathrm{M}_\Lambda, s \models^{\mathbf{1}} \psi_1$ and $\mathrm{M}_\Lambda, s \models^{\mathbf{1}} \psi_2$) iff $\psi_1 \in s$ and $\psi_2 \in s$ (by induction on (1)) iff $\psi_1 \wedge \psi_2 \in s$ (since s is a (maximal) theory).
- $\psi = \neg\psi_1$. Since $(0,\psi_1) < (0,\neg\psi_1)$ we have that $\mathrm{M}_\Lambda, s \models^{\mathbf{1}} \neg\psi_1$ iff not$(\mathrm{M}_\Lambda, s \models^{\mathbf{1}} \psi_1)$ iff not$(\psi_1 \in s)$ (by induction on (1)) iff $\neg\psi_1 \in s$ since s is (a) maximal (theory).
- $\psi = \mathbf{K}_i\psi_1$. We will prove two implications:
  '$\Leftarrow$' Suppose $\mathbf{K}_i\psi_1 \in s$. Then by definition of $\mathrm{R}_\Lambda(i)$, $\psi_1 \in t$ for all t such that $(s,t) \in \mathrm{R}_\Lambda(i)$. Since $(0,\psi_1) < (0,\mathbf{K}_i\psi_1)$, this implies that $\mathrm{M}_\Lambda, t \models^{\mathbf{1}} \psi_1$ for all $t \in \mathrm{S}_\Lambda$ with $(s,t) \in \mathrm{R}_\Lambda(i)$, hence $\mathrm{M}_\Lambda, s \models^{\mathbf{1}} \mathbf{K}_i\psi_1$.
  '$\Rightarrow$' Suppose $\mathrm{M}_\Lambda, s \models^{\mathbf{1}} \mathbf{K}_i\psi_1$. Now if for $t \in \mathrm{S}_\Lambda$, $(s,t) \in \mathrm{R}_\Lambda(i)$, then $\mathrm{M}_\Lambda, t \models^{\mathbf{1}} \psi_1$. Since $(0,\psi_1) < (0,\mathbf{K}_i\psi_1)$, we have by induction on (1) that $\psi_1 \in t$, for all $t \in \mathrm{S}_\Lambda$ with $(s,t) \in \mathrm{R}_\Lambda(i)$. This implies that ψ_1 belongs to every maximal theory containing $s/\mathbf{K}_i$, and by Proposition 3.45 we conclude that $s/\mathbf{K}_i \vdash^\Lambda \psi_1$. By Corollary 3.42 we conclude that $\mathbf{K}_i\psi_1 \in s$.
- $\psi = [\mathrm{do}_i(\alpha)]\psi_1$. We will prove two implications:
  '$\Leftarrow$' Let $[\mathrm{do}_i(\alpha)]\psi_1 \in s$. Let $t = \mathtt{r}^{\mathbf{1}}(i,\alpha)(s)$. Since $(1,[\mathrm{do}_i(\alpha)]\psi_1) < (0,[\mathrm{do}_i(\alpha)]\psi_1)$, we find by induction on (2b) that $\psi_1 \in t$. Since $(0,\psi_1) < (0,[\mathrm{do}_i(\alpha)]\psi_1)$, we find by induction on (1) that $\mathrm{M}_\Lambda, t \models^{\mathbf{1}} \psi_1$, if $t = \mathtt{r}^{\mathbf{1}}(i,\alpha)(s)$. But this implies that $\mathrm{M}_\Lambda, s \models^{\mathbf{1}} [\mathrm{do}_i(\alpha)]\psi_1$.
  '$\Rightarrow$' Suppose $\mathrm{M}_\Lambda, s \models^{\mathbf{1}} [\mathrm{do}_i(\alpha)]\psi_1$. This implies that $\mathrm{M}_\Lambda, t \models^{\mathbf{1}} \psi_1$ if $t = \mathtt{r}^{\mathbf{1}}(i,\alpha)(s)$. Since $(0,\psi_1) < (0,[\mathrm{do}_i(\alpha)]\psi_1)$ we have by induction on (1) that $\psi_1 \in t$ if $t = \mathtt{r}^{\mathbf{1}}(i,\alpha)(s)$. Now since $(1,[\mathrm{do}_i(\alpha)]\psi_1) < (0,[\mathrm{do}_i(\alpha)]\psi_1)$, we conclude by induction on (2a) that $[\mathrm{do}_i(\alpha)]\psi_1 \in s$.
- $\psi = \mathbf{A}_i\alpha$. Since $(1,\mathbf{A}_i\alpha) < (0,\mathbf{A}_i\alpha)$ we find by induction on (3) that $\mathrm{M}_\Lambda, s \models^{\mathbf{1}} \mathbf{A}_i\alpha$ iff $\mathtt{c}^{\mathbf{1}}(i,\alpha)(s) = \mathbf{1}$ iff $\mathbf{A}_i\alpha \in s$.

Next we prove (2a). Let $(1,[\mathrm{do}_i(\alpha)]\psi) < (0,\varphi)$. Consider the various possibilities for α.

- $\alpha = a$, for $a \in \mathrm{At}$. Assume that $\psi \in t$ if $t = \mathtt{r}_\Lambda(i,a)(s)$. By definition of $\mathtt{r}_\Lambda$ this implies that ψ is in every maximal theory containing $s/[\mathrm{do}_i(a)]$, i.e. $s/[\mathrm{do}_i(a)] \vdash^\Lambda \psi$.

By Corollary 3.42 we conclude that $[\mathrm{do}_i(a)]\psi \in s$.

- $\alpha = \mathtt{confirm}\,\psi_1$. Assume that $\psi \in t$ for $t = \mathtt{r^1}(i, \mathtt{confirm}\,\psi_1)(s)$. If $\mathrm{M}_\Lambda, s \models^{\mathbf{1}} \psi_1$ we have that $s = t$, by definition of $\mathtt{r^1}$. Then $\psi \in s$, and, since s is a theory, this implies $\neg\psi_1 \vee \psi \in s$, which on its turn implies $[\mathrm{do}_i(\mathtt{confirm}\,\psi_1)]\psi \in s$. If $\mathrm{M}_\Lambda, s \models^{\mathbf{1}} \neg\psi_1$ then, since $(0, \neg\psi_1) < (1, [\mathrm{do}_i(\mathtt{confirm}\,\psi_1)]\psi)$, we have by induction on (1) that $\neg\psi_1 \in s$, hence $\neg\psi_1 \vee \psi \in s$, and thus $[\mathrm{do}_i(\mathtt{confirm}\,\psi_1)]\psi \in s$.
- $\alpha = \alpha_1; \alpha_2$. Assume that $\psi \in t$ for $t = \mathtt{r^1}(i, \alpha_1; \alpha_2)(s)$. By definition of $\mathtt{r^1}$ this implies that $\psi \in t$ for $t = \mathtt{r^1}(i, \alpha_2)(u)$ for $u = \mathtt{r^1}(i, \alpha_1)(s)$. Since $(1, [\mathrm{do}_i(\alpha_2)]\psi) < (1, [\mathrm{do}_i(\alpha_1; \alpha_2)]\psi)$ we have that $[\mathrm{do}_i(\alpha_2)]\psi \in u$ for $u = \mathtt{r^1}(i, \alpha_1)(s)$. Since furthermore $(1, [\mathrm{do}_i(\alpha_1)][\mathrm{do}_i(\alpha_2)]\psi) < (1, [\mathrm{do}_i(\alpha_1; \alpha_2)]\psi)$ we have that $[\mathrm{do}_i(\alpha_1)][\mathrm{do}_i(\alpha_2)]\psi \in s$. Since s is closed under the axioms of $\Sigma_\mathbf{1}$ and MP, this implies that $[\mathrm{do}_i(\alpha_1; \alpha_2)]\psi \in s$.
- $\alpha = \mathtt{if}\,\varphi\,\mathtt{then}\,\alpha_1\,\mathtt{else}\,\alpha_2\,\mathtt{fi}$. Let $\psi \in t$ for $t = \mathtt{r^1}(i, \mathtt{if}\,\varphi\,\mathtt{then}\,\alpha_1\,\mathtt{else}\,\alpha_2\,\mathtt{fi})(s)$. Then $\psi \in t$ for all $t = \mathtt{r^1}(i, \mathtt{confirm}\,\varphi; \alpha_1)(s)$ and $\psi \in t$ for all $t = \mathtt{r^1}(i, \mathtt{confirm}\,\neg\varphi; \alpha_2)(s)$. Since we have both $(1, [\mathrm{do}_i(\mathtt{confirm}\,\varphi; \alpha_1)]\psi) < (1, [\mathrm{do}_i(\mathtt{if}\,\varphi\,\mathtt{then}\,\alpha_1\,\mathtt{else}\,\alpha_2\,\mathtt{fi})]\psi)$ and $(1, [\mathrm{do}_i(\mathtt{confirm}\,\neg\varphi; \alpha_2)]\psi) < (1, [\mathrm{do}_i(\mathtt{if}\,\varphi\,\mathtt{then}\,\alpha_1\,\mathtt{else}\,\alpha_2\,\mathtt{fi})]\psi)$ we have by induction that $[\mathrm{do}_i(\mathtt{confirm}\,\varphi; \alpha_1)]\psi \in s$ and $[\mathrm{do}_i(\mathtt{confirm}\,\neg\varphi; \alpha_2)]\psi \in s$, and, since s is a theory, $[\mathrm{do}_i(\mathtt{if}\,\varphi\,\mathtt{then}\,\alpha_1\,\mathtt{else}\,\alpha_2\,\mathtt{fi})]\psi \in s$.
- $\alpha = \mathtt{while}\,\varphi\,\mathtt{do}\,\alpha\,\mathtt{od}$. Assume that $\psi \in t$ for $t = \mathtt{r^1}(i, \mathtt{while}\,\varphi\,\mathtt{do}\,\alpha\,\mathtt{od})(s)$. Since $\mathtt{r^1}(i, \mathtt{while}\,\varphi\,\mathtt{do}\,\alpha\,\mathtt{od})(s) = \cup_{k\in\mathbb{N}}\mathtt{r^1}(i, (\mathtt{confirm}\,\varphi; \alpha)^k; \mathtt{confirm}\,\neg\varphi)(s)$, it follows that $\psi \in t$ for all $t = \mathtt{r^1}(i, (\mathtt{confirm}\,\varphi; \alpha)^k; \mathtt{confirm}\,\neg\varphi)(s)$, for all $k \in \mathbb{N}$. Now since $(1, [\mathrm{do}_i((\mathtt{confirm}\,\varphi; \alpha)^k; \mathtt{confirm}\,\neg\varphi)]\psi) < (1, [\mathrm{do}_i(\mathtt{while}\,\varphi\,\mathtt{do}\,\alpha\,\mathtt{od})]\psi)$ for all $k \in \mathbb{N}$ we have by induction on (2a) that $\psi_k(i, \varphi, \alpha) \in s$ for all $k \in \mathbb{N}$, and since s is closed under ΩA this implies that $[\mathrm{do}_i(\mathtt{while}\,\varphi\,\mathtt{do}\,\alpha\,\mathtt{od})]\psi \in s$.

We continue with proving (2b). So let again $(1, [\mathrm{do}_i(\alpha)]\psi) < (0, \varphi)$, and consider the various possibilities for α.

- $\alpha = a$, for $a \in \mathrm{At}$. If $t = \mathtt{r}_\Lambda(i, a)(s)$ and $[\mathrm{do}_i(a)]\psi \in s$, then by definition of $\mathtt{r}_\Lambda$, $\psi \in t$.
- $\alpha = \mathtt{confirm}\,\psi_1$. Let $t = \mathtt{r^1}(i, \mathtt{confirm}\,\psi_1)(s)$ and $[\mathrm{do}_i(\mathtt{confirm}\,\psi_1)]\psi \in s$. By definition of $\mathtt{r^1}$, $\mathrm{M}_\Lambda, s \models^{\mathbf{1}} \psi_1$ and $s = t$. Since $(0, \psi_1) < (0, \neg\psi_1) < (1, [\mathrm{do}_i(\mathtt{confirm}\,\psi_1)]\psi)$, we find by induction on (1) that $\psi_1 \in s$. Since s is a theory, $[\mathrm{do}_i(\mathtt{confirm}\,\psi_1)]\psi \in s$ implies that $\neg\psi_1 \vee \psi \in s$, and, since s is maximal, we conclude that $\psi \in s$.
- $\alpha = \alpha_1; \alpha_2$. Let $t = \mathtt{r^1}(i, \alpha_1; \alpha_2)(s)$ and $[\mathrm{do}_i(\alpha_1; \alpha_2)]\psi \in s$. Then, by definition of $\mathtt{r^1}$, we have that $t = \mathtt{r^1}(i, \alpha_2)(u)$ for some $u \in \mathrm{S}_\Lambda$ such that $u = \mathtt{r^1}(i, \alpha_1)(s)$. Since s is closed under the axioms and proof rules of $\Sigma_\mathbf{1}$ we have that $[\mathrm{do}_i(\alpha_1)][\mathrm{do}_i(\alpha_2)]\psi \in s$, and hence, since $(1, [\mathrm{do}_i(\alpha_1)][\mathrm{do}_i(\alpha_2)]\psi) < (1, [\mathrm{do}_i(\alpha_1; \alpha_2)]\psi)$, we have by induction on (2b) that $[\mathrm{do}_i(\alpha_2)]\psi \in u$. But this implies, since $(1, [\mathrm{do}_i(\alpha_2)]\psi) < (1, [\mathrm{do}_i(\alpha_1; \alpha_2)]\psi)$, that $\psi \in t$.
- $\alpha = \mathtt{if}\,\varphi\,\mathtt{then}\,\alpha_1\,\mathtt{else}\,\alpha_2\,\mathtt{fi}$. Let $t = \mathtt{r^1}(i, \mathtt{if}\,\varphi\,\mathtt{then}\,\alpha_1\,\mathtt{else}\,\alpha_2\,\mathtt{fi})(s)$ and let furthermore $[\mathrm{do}_i(\mathtt{if}\,\varphi\,\mathtt{then}\,\alpha_1\,\mathtt{else}\,\alpha_2\,\mathtt{fi})]\psi \in s$. Then either $t = \mathtt{r^1}(i, \mathtt{confirm}\,\varphi; \alpha_1)(s)$ or

$t = \mathtt{r}^{\mathbf{1}}(i, \mathtt{confirm}\,\neg\varphi; \alpha_2)(s)$. If $[\mathrm{do}_i(\mathtt{if}\,\varphi\,\mathtt{then}\,\alpha_1\,\mathtt{else}\,\alpha_2\,\mathtt{fi})]\psi \in s$,then, since s is a theory, both $[\mathrm{do}_i(\mathtt{confirm}\,\varphi; \alpha_1)]\psi \in s$ and $[\mathrm{do}_i(\mathtt{confirm}\,\neg\varphi; \alpha_2)]\psi \in s$. Since it holds that both $(1, [\mathrm{do}_i(\mathtt{confirm}\,\varphi; \alpha_1)]\psi) < (1, [\mathrm{do}_i(\mathtt{if}\,\varphi\,\mathtt{then}\,\alpha_1\,\mathtt{else}\,\alpha_2\,\mathtt{fi})]\psi)$ and $(1, [\mathrm{do}_i(\mathtt{confirm}\,\neg\varphi; \alpha_2)]\psi) < (1, [\mathrm{do}_i(\mathtt{if}\,\varphi\,\mathtt{then}\,\alpha_1\,\mathtt{else}\,\alpha_2\,\mathtt{fi})]\psi)$ we have by induction on (2b) that $\psi \in t$.

- $\alpha = \mathtt{while}\,\varphi\,\mathtt{do}\,\alpha\,\mathtt{od}$. Let $t = \mathtt{r}^{\mathbf{1}}(i, \mathtt{while}\,\varphi\,\mathtt{do}\,\alpha\,\mathtt{od})(s)$ and $[\mathrm{do}_i(\mathtt{while}\,\varphi\,\mathtt{do}\,\alpha\,\mathtt{od})]\psi \in s$. Since s is a theory, we have that $\psi_l(i, \varphi, \alpha) \in s$, for all $l \in \mathbb{N}$. By definition of $\mathtt{r}^{\mathbf{1}}$, it holds that $t = \mathtt{r}^{\mathbf{1}}(i, (\mathtt{confirm}\,\varphi; \alpha)^k; \mathtt{confirm}\,\neg\varphi)(s)$ for some $k \in \mathbb{N}$. Now since $(1, \psi_l(i, \varphi, \alpha)) < (1, [\mathrm{do}_i(\mathtt{while}\,\varphi\,\mathtt{do}\,\alpha\,\mathtt{od})]\psi)$ for all $l \in \mathbb{N}$, we conclude by induction on (2b) that $\psi \in t$.

Finally we come to the proof of item (3). Let $(1, \mathbf{A}_i\alpha) < (0, \varphi)$. Consider the various cases for α.

- $\alpha = a$, where $a \in \mathrm{At}$. Now $\mathtt{c}_\Lambda(i, a)(s) = \mathbf{1}$ iff $\mathbf{A}_i a \in s$, by definition of $\mathtt{c}_\Lambda$.
- $\alpha = \mathtt{confirm}\,\psi_1$. By definition, $\mathtt{c}^{\mathbf{1}}(i, \mathtt{confirm}\,\psi_1)(s) = \mathbf{1}$ iff $\mathrm{M}_\Lambda, s \models^{\mathbf{1}} \psi_1$ iff, since $(0, \psi_1) < (0, \mathbf{A}_i\mathtt{confirm}\,\psi_1)$, $\psi_1 \in s$ iff $\mathbf{A}_i\mathtt{confirm}\,\psi_1$ in s, since s is a theory.
- $\alpha = \alpha_1; \alpha_2$. We prove two implications:

  '$\Leftarrow$' Since s is a theory, $\mathbf{A}_i\alpha_1; \alpha_2 \in s$ iff $\mathbf{A}_i\alpha_1 \in s$ and $[\mathrm{do}_i(\alpha_1)]\mathbf{A}_i\alpha_2 \in s$. Since $(1, \mathbf{A}_i\alpha_1) < (1, \mathbf{A}_i\alpha_1; \alpha_2)$, we find by induction on (3) that $\mathtt{c}^{\mathbf{1}}(i, \alpha_1)(s) = \mathbf{1}$. Now suppose $t = \mathtt{r}^{\mathbf{1}}(i, \alpha_1)(s)$. Since $(1, [\mathrm{do}_i(\alpha_1)]\mathbf{A}_i\alpha_2) < (1, \mathbf{A}_i\alpha_1; \alpha_2)$, we find by induction on (2b) that $\mathbf{A}_i\alpha_2 \in t$. Furthermore, since $(1, \mathbf{A}_i\alpha_2) < (1, \mathbf{A}_i\alpha_1; \alpha_2)$, the latter implies that $\mathtt{c}^{\mathbf{1}}(i, \alpha_2)(t) = \mathbf{1}$, for all $t = \mathtt{r}^{\mathbf{1}}(i, \alpha_1)(s)$, which, together with $\mathtt{c}^{\mathbf{1}}(i, \alpha_1)(s) = \mathbf{1}$, suffices to conclude that $\mathtt{c}^{\mathbf{1}}(i, \alpha_1; \alpha_2)(s) = \mathbf{1}$.

  '$\Rightarrow$' By definition, $\mathtt{c}^{\mathbf{1}}(i, \alpha_1; \alpha_2)(s) = \mathbf{1}$ iff $\mathtt{c}^{\mathbf{1}}(i, \alpha_1)(s) = \mathbf{1}$ and $\mathtt{c}^{\mathbf{1}}(i, \alpha_2)(t) = \mathbf{1}$ for all $t = \mathtt{r}^{\mathbf{1}}(i, \alpha_1)(s)$. Now since $(1, \mathbf{A}_i\alpha_1) < (1, \mathbf{A}_i\alpha_1; \alpha_2)$, we conclude by induction on (3) that $\mathbf{A}_i\alpha_1 \in s$. Furthermore, since $(1, \mathbf{A}_i\alpha_2) < (1, \mathbf{A}_i\alpha_1; \alpha_2)$, we have that $\mathbf{A}_i\alpha_2 \in t$, for all $t = \mathtt{r}^{\mathbf{1}}(i, \alpha_1)(s)$. Now since $(1, [\mathrm{do}_i(\alpha_1)]\mathbf{A}_i\alpha_2) < (1, \mathbf{A}_i\alpha_1; \alpha_2)$, we find by induction on (2a) that $[\mathrm{do}_i(\alpha_1)]\mathbf{A}_i\alpha_2 \in s$. But then, since s is a theory, we conclude that $\mathbf{A}_i\alpha_1; \alpha_2 \in s$.
- $\alpha = \mathtt{if}\,\varphi\,\mathtt{then}\,\alpha_1\,\mathtt{else}\,\alpha_2\,\mathtt{fi}$. By definition of $<$ we have that $(1, \mathbf{A}_i\mathtt{confirm}\,\varphi; \alpha_1) < (1, \mathbf{A}_i\mathtt{if}\,\varphi\,\mathtt{then}\,\alpha_1\,\mathtt{else}\,\alpha_2\,\mathtt{fi})$ and furthermore that $(1, \mathbf{A}_i\mathtt{confirm}\,\neg\varphi; \alpha_2) < (1, \mathbf{A}_i\mathtt{if}\,\varphi\,\mathtt{then}\,\alpha_1\,\mathtt{else}\,\alpha_2\,\mathtt{fi})$. This implies that $\mathtt{c}^{\mathbf{1}}(i, \mathtt{if}\,\varphi\,\mathtt{then}\,\alpha_1\,\mathtt{else}\,\alpha_2\,\mathtt{fi})(s) = \mathbf{1}$ iff $\mathtt{c}^{\mathbf{1}}(i, \mathtt{confirm}\,\varphi; \alpha_1)(s) = \mathbf{1}$ or $\mathtt{c}^{\mathbf{1}}(i, \mathtt{confirm}\,\neg\varphi; \alpha_2)(s) = \mathbf{1}$ iff — by induction on (3) — $\mathbf{A}_i\mathtt{confirm}\,\varphi; \alpha_1 \in s$ or $\mathbf{A}_i\mathtt{confirm}\,\neg\varphi; \alpha_2 \in s$ iff $\mathbf{A}_i\mathtt{if}\,\varphi\,\mathtt{then}\,\alpha_1\,\mathtt{else}\,\alpha_2\,\mathtt{fi} \in s$, since s is a theory.
- $\alpha = \mathtt{while}\,\varphi\,\mathtt{do}\,\alpha\,\mathtt{od}$. We prove two implications:

  '$\Leftarrow$' Let $\mathbf{A}_i\mathtt{while}\,\varphi\,\mathtt{do}\,\alpha\,\mathtt{od} \in s$. Then, since s is maximal, $\neg\mathbf{A}_i\mathtt{while}\,\varphi\,\mathtt{do}\,\alpha\,\mathtt{od} \notin s$, and, since s is closed under $\Omega\mathrm{IA}$, this implies that $\neg(\varphi_k(i, \alpha)) \notin s$, for some $k \in \mathbb{N}$, and, again due to the maximality of s, $\varphi_k(i, \alpha) \in s$. Since $(1, \varphi_l(i, \alpha)) <$

$(1, \mathbf{A}_i\texttt{while}\,\varphi\,\texttt{do}\,\alpha\,\texttt{od})$ for all $l \in \mathbb{N}$, we have by induction on (3) that $\mathsf{c}^{\mathbf{1}}(i, (\texttt{confirm}\,\varphi; \alpha)^k; \texttt{confirm}\,\neg\varphi)(s) = \mathbf{1}$, and, by definition of $\mathsf{c}^{\mathbf{1}}$, this implies $\mathsf{c}^{\mathbf{1}}(i, \texttt{while}\,\varphi\,\texttt{do}\,\alpha\,\texttt{od})(s) = \mathbf{1}$.

'$\Rightarrow$' If $\mathsf{c}^{\mathbf{1}}(i, \texttt{while}\,\varphi\,\texttt{do}\,\alpha\,\texttt{od})(s) = \mathbf{1}$, then $\mathsf{c}^{\mathbf{1}}(i, (\texttt{confirm}\,\varphi; \alpha_1)^k; \texttt{confirm}\,\neg\varphi)(s) = \mathbf{1}$ for some $k \in \mathbb{N}$. Since $(1, \varphi_l(i, \alpha)) < (1, \mathbf{A}_i\texttt{while}\,\varphi\,\texttt{do}\,\alpha\,\texttt{od})$ for all $l \in \mathbb{N}$, this implies by induction on (3) that $\varphi_k(i, \alpha) \in s$. Then, since s is a theory, $\neg(\varphi_k(i, \alpha)) \notin s$, and, by item 7 of Proposition 3.34, it follows that $\neg\mathbf{A}_i\texttt{while}\,\varphi\,\texttt{do}\,\alpha\,\texttt{od} \notin s$. Now since s is maximal it follows that $\mathbf{A}_i\texttt{while}\,\varphi\,\texttt{do}\,\alpha\,\texttt{od} \in s$.

Having proved the items (1), (2) and (3) suffices to conclude that the truth-theorem holds.

⊠

The proof of the truth-theorem for **0**-logics is almost identical to the one given for Theorem 3.54. One just needs to change one clause in the definition of the $\prec$-relation, used to apply induction upon, and modify the proof of the truth-theorem accordingly.

3.55. Definition. The ordering $<'$ is defined as the transitive closure of the smallest relation on $\{0, 1\} \times \mathrm{L}$ satisfying the constraints 1 through 15 and 17 through 20 as given in Definition 3.51 and the constraint

16.' $(1, [\mathrm{do}_i(\alpha_1)]\neg\mathbf{A}_i\alpha_2) \prec (1, \mathbf{A}_i\alpha_1; \alpha_2)$

The ordering $\leq'$ is defined to be the reflexive closure of $<'$.

The only modification to the proof of the truth-theorem for **1**-logics that is sufficient to end up with a proof of a truth-theorem for **0**-logics concerns the proof of property (3) for sequentially composed actions, i.e. the proof that $\mathsf{c}^{\mathbf{0}}(i, \alpha_1; \alpha_2)(s) = \mathbf{1}$ iff $\mathbf{A}_i\alpha_1; \alpha_2 \in s$, whenever $(1, \mathbf{A}_i\alpha_1; \alpha_2) <' (0, \varphi)$. We will show this by proving two implications:

'$\Leftarrow$' Since s is a Λ-theory, $\mathbf{A}_i\alpha_1; \alpha_2$ in s iff $\mathbf{A}_i\alpha_1 \in s$ and $\neg[\mathrm{do}_i(\alpha_1)]\neg\mathbf{A}_i\alpha_2 \in s$. Since $(1, \mathbf{A}_i\alpha_1) <' (1, \mathbf{A}_i\alpha_1; \alpha_2)$ we find by induction on (3) that $\mathsf{c}^{\mathbf{0}}(i, \alpha_1)(s) = \mathbf{1}$. Since $(1, [\mathrm{do}_i(\alpha_1)]\neg\mathbf{A}_i\alpha_2) <' (1, \mathbf{A}_i\alpha_1; \alpha_2)$, we find by induction on (2b), read in its contrapositive form, that for some $t \in \mathrm{S}_\Lambda$, $t = \mathsf{r}^{\mathbf{0}}(i, \alpha_1)(s)$ with $\neg\mathbf{A}_i\alpha_2 \notin t$. Now since t is maximal, this implies that $\mathbf{A}_i\alpha_2 \in t$. Since $(1, \mathbf{A}_i\alpha_2) <' (1, \mathbf{A}_i\alpha_1; \alpha_2)$ the latter implies that $\mathsf{c}^{\mathbf{0}}(i, \alpha_2)(t) = \mathbf{1}$. Together with $\mathsf{c}^{\mathbf{0}}(i, \alpha_1)(s) = \mathbf{1}$ this suffices to conclude that $\mathsf{c}^{\mathbf{0}}(i, \alpha_1; \alpha_2)(s) = \mathbf{1}$.

'$\Rightarrow$' By definition, $\mathsf{c}^{\mathbf{0}}(i, \alpha_1; \alpha_2)(s) = \mathbf{1}$ iff $\mathsf{c}^{\mathbf{0}}(i, \alpha_1)(s) = \mathbf{1}$ and for some $t \in \mathrm{S}_\Lambda$, $t = \mathsf{r}^{\mathbf{0}}(i, \alpha_1)(s)$ and $\mathsf{c}^{\mathbf{0}}(i, \alpha_2)(t) = \mathbf{1}$. Now since $(1, \mathbf{A}_i\alpha_1) <' (1, \mathbf{A}_i\alpha_1; \alpha_2)$ we have by induction on (3) that $\mathbf{A}_i\alpha_1 \in s$. Furthermore, since $(1, \mathbf{A}_i\alpha_2) <' (1, \mathbf{A}_i\alpha_1; \alpha_2)$, we have for the aforementioned t that $\mathbf{A}_i\alpha_2 \in t$. Hence we have some $t \in \mathrm{S}_\Lambda$ such that $t = \mathsf{r}^{\mathbf{0}}(i, \alpha_1)(s)$ and $\mathbf{A}_i\alpha_2 \in t$ while also $\mathbf{A}_i\alpha_1 \in s$. Since t is maximal, $\neg\mathbf{A}_i\alpha_2 \notin t$. And, rephrasing (2b) to 'if $t = \mathsf{r}^{\mathbf{0}}(i, \alpha)(s)$ and $\psi \notin t$ then $[\mathrm{do}_i(\alpha)]\psi \notin s$', we conclude by induction on (2b) that $[\mathrm{do}_i(\alpha_1)]\neg\mathbf{A}_i\alpha_2 \notin s$. Since s is maximal it follows that

$\neg[\mathrm{do}_i(\alpha_1)]\neg\mathbf{A}_i\alpha_2 \in s$. Hence $\mathbf{A}_i\alpha_1 \in s$ and $\langle\mathrm{do}_i(\alpha_1)\rangle\mathbf{A}_i\alpha_2 \in s$, which, since s is a Λ-theory, implies that $\mathbf{A}_i\alpha_1;\alpha_2 \in s$, which was to be shown.

Having proved the truth-theorem both for **1**-logics and for **0**-logics, we can prove that deducibility for a logic Λ corresponds with validity in the canonical model M_Λ.

3.56. Proposition. *For all* $\mathbf{b}$*-logics* Λ *and all* $\varphi \in \mathrm{L}$ *we have:* $\vdash^\Lambda \varphi$ *iff* $\mathrm{M}_\Lambda \models^{\mathbf{b}} \varphi$.

Proof: Let $\varphi \in \mathrm{L}$ be arbitrary. Then we have:

$\vdash^\Lambda \varphi$ iff $\varphi \in s$, for all $s \in \mathrm{S}_\Lambda$
iff $\mathrm{M}_\Lambda, s \models^{\mathbf{b}} \varphi$ for all $s \in \mathrm{S}_\Lambda$
iff $\mathrm{M}_\Lambda \models^{\mathbf{b}} \varphi$

⊠

Using the propositions and theorems shown above, we can now prove those given in Section 3.4. Note that Proposition 3.29 is already shown as the third item of Proposition 3.37.

3.26. Theorem. *For* $\mathbf{b} \in$ bool *and all* $\varphi \in \mathrm{L}$ *we have:*
$\vdash^{\mathbf{b}} \varphi \Leftrightarrow \models^{\mathbf{b}} \varphi$

Proof: We prove the theorem by proving two implications.
'$\Leftarrow$' If $\models^{\mathbf{b}} \varphi$ then $\mathrm{M} \models^{\mathbf{b}} \varphi$ for all $\mathrm{M} \in \mathbf{M}$. Since $\mathrm{M}_{\mathrm{LCap}_{\mathbf{b}}} \in \mathbf{M}$ it follows that $\mathrm{M}_{\mathrm{LCap}_{\mathbf{b}}} \models^{\mathbf{b}} \varphi$. By Proposition 3.56 it then follows that $\vdash^{\mathbf{b}} \varphi$.
'$\Rightarrow$' Suppose $\vdash^{\mathbf{b}} \varphi$ and let $\mathrm{M} \in \mathbf{M}$. By Proposition 3.33 we have that $\{\psi \in \mathrm{L} \mid \mathrm{M} \models^{\mathbf{b}} \psi\}$ is a $\mathbf{b}$-logic. Since $\mathrm{LCap}_{\mathbf{b}}$ is the smallest $\mathbf{b}$-logic, it follows that whenever $\varphi \in \mathrm{LCap}_{\mathbf{b}}$ also $\varphi \in \{\psi \in \mathrm{L} \mid \mathrm{M} \models^{\mathbf{b}} \psi\}$, and hence $\mathrm{M} \models^{\mathbf{b}} \varphi$. Since M is arbitrary, it follows that $\mathrm{M} \models^{\mathbf{b}} \varphi$ for all $\mathrm{M} \in \mathbf{M}$ and thus $\models^{\mathbf{b}} \varphi$, which was to be shown.

⊠

3.30. Proposition. *The proof systems* $\Sigma_{\mathbf{1}}$ *and* $\Sigma_{\mathbf{0}}$ *are strongly complete, i.e. every set* $\Phi \subseteq \mathrm{L}$ *that is* $\mathrm{LCap}_{\mathbf{b}}$*-consistent is* $\models^{\mathbf{b}}$*-satisfiable.*

Proof: The proposition follows, for arbitrary logics, directly from the proof of Proposition 3.45. For if Φ is Λ-consistent, then by the procedure given in the proof of Proposition 3.45 one constructs a maximal Λ-theory Γ that contains Φ. This Γ appears as a state in the canonical model for Λ, and by the truth-theorems, all formulae from Γ — and hence from Φ — are satisfied at this state. Hence every Λ-consistent set Φ is satisfied at some state of the canonical model for Λ, and since this canonical model is a well-defined one, the proposition follows.

⊠

Chapter 4

Unravelling nondeterminism

The grey of evening fills the room,
There's no need to look outside,
To see or feel the rain.

Genesis, *'In The Glow Of The Light'*.

In this chapter we define various formal systems in which nondeterministic action constructors are considered. We deal with both internal and external nondeterminism. In the first kind of nondeterminism the agent makes the choice as to which action to take, in the latter some unspecified external environment does. It is assumed that the agent displays an *angelic* behaviour, whereas the external environment may display a *demonic* one. This means that the agent itself chooses the correct, desired action whenever possible, while it has to expect the external environment, which is completely unpredictable from the agent's point of view, to pick the incorrect, undesirable one if given that possibility. The presence of abilities in our system forces one to reconsider the common approaches towards nondeterminism. It turns out that the combination of internal nondeterminism and abilities cannot be formalised using (obvious extensions of) standard approaches. In the case of external nondeterminism one has to distinguish between optimistic and pessimistic agents as described in the previous chapter. For optimistic agents it is not possible to use any of the well-known approaches towards external nondeterminism (or concurrency), whereas for pessimistic agents an extension of the semantics for Concurrent Propositional Dynamic Logic as proposed by Peleg [101] may be used. For internal nondeterminism and external nondeterminism in the presence of optimistic agents we pursue an approach in which the (nondeterministic) actions are unravelled into the sequences of atomic actions and confirmations that constitute them. Based on this unravelling we present separate semantics for both internal and external nondeterminism that formalise an intuitively acceptable behaviour of agents with nondeterministic actions at their disposal. In particular, the validities describing the agents' abilities for

composite actions, containing nondeterministic action constructors, meet all intuitive requirements. It turns out that the Can-predicate and the Cannot-predicate, as formalising known practical (im)possibility, have to be reconsidered in the light of both internal and external nondeterminism, which leads to new definitions of these predicates in some cases. As usual, we conclude this chapter with a brief summary, ideas for possible extensions, references to the relevant literature, and proofs of selected propositions.

4.1 Internal versus external nondeterminism

Terms referring to nondeterministic actions are frequently used in common parlance. Statements like 'to see or feel the rain', 'take it or leave it, I don't care' or 'mail the letter or burn it, do as you like' (implicitly) refer to nondeterministic actions. In essence, nondeterminism of an action accounts to nothing but the outcome of execution of the action not being determined, i.e. execution of a nondeterministic action may have more than one possible outcome. The nondeterminism that we consider here originates from combining two actions into a composite action, representing the nondeterministic choice between the two actions. As remarked by Hoare [50], this choice can be made in two different ways: either the agent itself makes the choice, or something else does. In general we think of this 'something else' as some unspecified external environment, surrounding the agent but strictly separated from it. Dependent on who makes the choice, i.e. the agent or the external environment, two essentially different nondeterministic actions result. Following the terminology of Meyer [93], we refer to actions in which the nondeterministic choice is to be made by the agent as being *internal*, while actions in which the external environment chooses are termed *external*. We assume the agent to act in its own interest. This means that in situations in which a nondeterministic action can be performed both in a 'right' and in a 'wrong' way, the agent will choose to perform the action in the right way. Therefore the internal nondeterministic action displays a so-called *angelic* (cf. [11]) behaviour. The external environment on the other hand is completely out of the control of the agent, and therefore completely unpredictable from the agent's perspective. In particular, this implies that whenever there is the possibility of making a wrong choice, the agent has to take into account that the external environment makes this wrong choice. As such, from the agent's point of view, the behaviour of the external environment is to be considered *demonic* (again cf. [11]). This difference between angelic and demonic behaviour will be visible throughout the definitions formalising the internal and the external nondeterministic choice.

At this point it is important to contemplate on the meaning of the non-core formulae $\langle \mathrm{do}_i(\alpha) \rangle \varphi$ and $\mathbf{A}_i \alpha$ in the presence of nondeterministic actions. For the deterministic actions built up using the core action constructors, $\langle \mathrm{do}_i(\alpha) \rangle \varphi$ was intuitively interpreted as stating that agent i has the opportunity to do action α, and that φ result from i's exe-

cution of α. Formulae $\mathbf{A}_i\alpha$ were interpreted as stating that agent i has the reliable ability to perform action α. The intuitive interpretation of $\mathbf{A}_i\alpha$ takes on also for nondeterministic actions, where it is important to stress that the agent's ability has to be *reliable*, i.e. no matter how the opportunity presents itself, agent i is able to take it. The intuitive interpretation of $\langle \mathrm{do}_i(\alpha)\rangle\varphi$ that we propose for possibly nondeterministic actions α, is that agent i has the opportunity to perform α in such a way that φ is certain to result from this performance, or phrased differently, that agent i has the *reliable* opportunity to bring about φ by executing α. From this point of view the intuitive interpretation of the dual $[\mathrm{do}_i(\alpha)]\varphi$ is that the agent does not have the reliable opportunity to perform α in such a way that φ is avoided, or phrased differently, that every possible way of performing α open to the agent, i.e. for which the agent has the opportunity, results in φ being true. Note that for deterministic actions this interpretation coincides with the one used in Chapter 2.

4.2 Internal nondeterminism

As in the representation of Meyer [93], we denote the internal nondeterministic combination of two actions α_1 and α_2 by $\alpha_1 \oplus \alpha_2$. The idea is that execution of the action $\alpha_1 \oplus \alpha_2$ corresponds to execution of either one of α_1 or α_2, where the choice is up to the agent. Besides the action constructor $\oplus$, the language $\mathrm{L}^{\oplus}$ furthermore contains primitive operators representing the Can-predicate and the Cannot-predicate. The reasons for declaring these predicates primitive, rather than introducing them through abbreviation as we did in the previous chapter, will be made clear later on.

4.1. Definition. To define the language $\mathrm{L}^{\oplus}$, the alphabet is extended with the action constructor $\_ \oplus \_$, representing the internal nondeterministic choice operator, and the predicates $\mathbf{Can}\_(\_,\_)$ and $\mathbf{Cannot}\_(\_,\_)$. The language $\mathrm{L}^{\oplus}$ is the smallest superset of Π closed under the core clauses. The class $\mathrm{Ac}^{\oplus}$ of actions is the smallest superset of At closed under the core clauses and such that $\alpha_1 \oplus \alpha_2 \in \mathrm{Ac}^{\oplus}$ whenever $\alpha_1 \in \mathrm{Ac}^{\oplus}, \alpha_2 \in \mathrm{Ac}^{\oplus}$.

It is obvious that the language $\mathrm{L}^{\oplus}$ as defined above is a genuine extension of the language L as defined in the previous chapter.

As for the core action constructors, we have to decide how to interpret and define the ability, opportunity and result for internal nondeterministic actions. Given the fact that the agent, which decides what action to take, displays an angelic behaviour, it seems reasonable to declare the agent to have the reliable opportunity to perform $\alpha_1 \oplus \alpha_2$ in such a way that φ results iff it either has the reliable opportunity to do α_1 such that φ results or it has the reliable opportunity to perform α_2 in such a way that execution leads to φ. From a formal point of view this corresponds to the equivalence

$$\langle \mathrm{do}_i(\alpha_1 \oplus \alpha_2)\rangle\varphi \leftrightarrow \langle \mathrm{do}_i(\alpha_1)\rangle\varphi \vee \langle \mathrm{do}_i(\alpha_2)\rangle\varphi \qquad (\dagger)$$

being valid. With regard to the ability of an agent i to perform the action $\alpha_1 \oplus \alpha_2$ an analogous argument as for the opportunity can be given. Since the agent itself chooses, and it acts in its own interests, it suffices that the agent is reliably able to perform either α_1 or α_2 in order to conclude that it is reliably able to perform $\alpha_1 \oplus \alpha_2$. Conversely, it is also necessary that the agent has the reliable ability to perform at least one of α_1 and α_2 for it to be reliably able to perform $\alpha_1 \oplus \alpha_2$, since it has to choose one of these actions. Formally this amounts to the formula

$$\mathbf{A}_i(\alpha_1 \oplus \alpha_2) \leftrightarrow \mathbf{A}_i\alpha_1 \vee \mathbf{A}_i\alpha_2 \qquad (\ddagger)$$

being valid. Accepting these equivalences to completely describe the behaviour of $\alpha_1 \oplus \alpha_2$, one could attempt to straightforwardly incorporate these in the framework of the previous chapter. However, this straightforward approach leads to undesirable, i.e. intuitively unacceptable, situations for the agent's ability with respect to sequentially composed actions. For intuitively one expects the ability to perform an action $(\alpha_1 \oplus \alpha_2);\alpha_3$ to be equivalent to either having the ability to do $\alpha_1;\alpha_3$ or having the ability to perform $\alpha_2;\alpha_3$. This equivalence does however not come about when pursuing the straightforward approach, as can be seen in the following example.

4.2. Example. Assume an agent i to be in a state such that $\mathbf{A}_i a_1, \neg\mathbf{A}_i a_2, \langle \mathrm{do}_i(a_1)\rangle\neg\mathbf{A}_i a_3$ and $\langle \mathrm{do}_i(a_2)\rangle\mathbf{A}_i a_3$ are true in the state, for $a_1, a_2, a_3 \in \mathrm{At}$. That is, the agent has both the ability and the opportunity to do a_1, and the opportunity but not the ability to do a_2, where doing a_2 results in i having the ability to do a_3 while doing a_1 does not. Both in the optimistic and in the pessimistic approach of the previous chapter, i is neither able to do $a_1;a_3$ nor to do $a_2;a_3$. However, using the equivalences $(\dagger)$ and $(\ddagger)$ as given above, one has to conclude that the agent is able to do $(a_1 \oplus a_2);a_3$. For since $\mathbf{A}_i a_1$ is true, $\mathbf{A}_i(a_1 \oplus a_2)$ is also true, according to $(\ddagger)$. Moreover, using $(\dagger)$ it follows that $\langle \mathrm{do}_i(a_2)\rangle\mathbf{A}_i a_3$ implies that $\langle \mathrm{do}_i(a_1 \oplus a_2)\rangle\mathbf{A}_i a_3$ holds, which both in the optimistic and in the pessimistic approach of the previous chapter implies that $\mathbf{A}_i((a_1 \oplus a_2);a_3)$ holds. Hence even though the agent is neither able to do $a_1;a_3$ nor to do $a_2;a_3$, it is able to do $\mathbf{A}_i((a_1 \oplus a_2);a_3)$, a conclusion which is clearly unacceptable.

Example 4.2 shows that the straightforward approach, embodied in the combination of the equivalences $(\dagger)$ and $(\ddagger)$ and the equivalences for sequential composition derived in the previous chapter, is not suited to capture the intuition that the action $(a_1 \oplus a_2);a_3$ is somehow composed of the actions $a_1;a_3$ and $a_2;a_3$. To solve this clash between formalism and intuition, we propose an approach in which an action is equated with the more elementary actions that constitute it. In the case of Example 4.2 this implies that the action $(a_1 \oplus a_2);a_3$ is also *formally* — and not just *intuitively* — equated with

the combination of $a_1;a_3$ and $a_2;a_3$. In particular, it is the case that an agent is able to perform the action $(a_1 \oplus a_2);a_3$ if and only if it is either able to perform $a_1;a_3$ or it is able to perform $a_2;a_3$. To formalise this idea of decomposing actions into their more elementary constituents, we use the unravel function $\mathrm{CS}^{\oplus}$, already discussed in Chapter 2. Using the function $\mathrm{CS}^{\oplus}$ we have the possibility to look at actions from a reductionistic point of view, i.e. in terms of their elementary constituents, the finite computation sequences.

4.3. Definition. The function $\mathrm{CS}^{\oplus} : \mathrm{Ac}^{\oplus} \to \wp(\mathrm{Ac}^{\oplus}_{\mathrm{b}})$ is for the non-core action constructor $\oplus$ defined by: $\mathrm{CS}^{\oplus}(\alpha_1 \oplus \alpha_2) = \mathrm{CS}^{\oplus}(\alpha_1) \cup \mathrm{CS}^{\oplus}(\alpha_2)$.

The presence of nondeterministic action constructors in the language does not necessitate the incorporation of new semantic constructs in the models for $\mathrm{L}^{\oplus}$: these models consist of nothing but the core elements.

4.4. Definition. A model M for the language $\mathrm{L}^{\oplus}$ is a tuple consisting of the core elements. The class of all models for $\mathrm{L}^{\oplus}$ is denoted by $\mathbf{M}^{\oplus}$.

The fundamental idea underlying the definition of $\models^{\oplus}$ is to treat an action as it were the existentially quantified union of all its finite computation sequences.

4.5. Definition. The binary relation $\models^{\oplus}$ between a formula from $\mathrm{L}^{\oplus}$ and a pair M, s consisting of a model M for $\mathrm{L}^{\oplus}$ and a state s in M is for the non-core formulae defined as follows:

$$\begin{array}{ll}
\mathrm{M}, s \models^{\oplus} \langle \mathrm{do}_i(\alpha)\rangle\varphi & \Leftrightarrow \exists \alpha' \in \mathrm{CS}^{\oplus}(\alpha) \exists s' \in \mathrm{S}(\mathtt{r}^{\oplus}(i,\alpha')(s) = s' \,\&\, \mathrm{M}, s' \models^{\oplus} \varphi) \\
\mathrm{M}, s \models^{\oplus} \mathbf{A}_i\alpha & \Leftrightarrow \exists \alpha' \in \mathrm{CS}^{\oplus}(\alpha)(\mathtt{c}^{\oplus}(i,\alpha')(s) = \mathbf{1})
\end{array}$$

where $\mathtt{r}^{\oplus}$ and $\mathtt{c}^{\oplus}$ are defined by:

$$\begin{array}{lll}
\mathtt{r}^{\oplus} & : & \mathrm{A} \times \mathrm{Ac}^{\oplus}_{\mathrm{b}} \to \mathrm{S}^{\cdot} \to \mathrm{S}^{\cdot} \\
\mathtt{r}^{\oplus}(i,a)(s) & = & \mathtt{r}_{\mathrm{o}}(i,a)(s) \\
\mathtt{r}^{\oplus}(i,\mathtt{confirm}\,\varphi)(s) & = & s \text{ if } \mathrm{M}, s \models^{\oplus} \varphi \\
 & = & \emptyset \text{ otherwise} \\
\mathtt{r}^{\oplus}(i,\alpha_1;\alpha_2)(s) & = & \mathtt{r}^{\oplus}(i,\alpha_2)(\mathtt{r}^{\oplus}(i,\alpha_1)(s)) \\
\mathtt{r}^{\oplus}(i,\alpha)(\emptyset) & = & \emptyset
\end{array}$$

$$\begin{array}{lll}
\mathtt{c}^{\oplus} & : & \mathrm{A} \times \mathrm{Ac}^{\oplus}_{\mathrm{b}} \to \mathrm{S}^{\cdot} \to \mathsf{bool} \\
\mathtt{c}^{\oplus}(i,a)(s) & = & \mathtt{c}_{\mathrm{o}}(i,a)(s) \\
\mathtt{c}^{\oplus}(i,\mathtt{confirm}\,\varphi)(s) & = & \mathbf{1} \text{ iff } \mathrm{M}, s \models^{\oplus} \varphi \\
\mathtt{c}^{\oplus}(i,\alpha_1;\alpha_2)(s) & = & \mathbf{1} \text{ iff } \mathtt{c}^{\oplus}(i,\alpha_1)(s) = \mathbf{1} \,\&\, \mathtt{c}^{\oplus}(i,\alpha_2)(\mathtt{r}^{\oplus}(i,\alpha_1)(s)) = \mathbf{1} \\
\mathtt{c}^{\oplus}(i,\alpha)(\emptyset) & = & \mathbf{1}
\end{array}$$

4.6. Remark. In Definition 4.5 we formalise only the optimistic approach towards the agents' abilities for sequentially composed actions. By replacing $\mathrm{c}^{\oplus}(i,\alpha)(\emptyset) = \mathbf{1}$ by $\mathrm{c}^{\oplus}(i,\alpha)(\emptyset) = \mathbf{0}$ one directly formalises the pessimistic approach, i.e. these approaches are completely orthogonal. In light of this orthogonality, we have decided to restrict ourselves to optimistic agents.

According to Definition 4.5, an agent has the ability to perform some action α iff it has the ability to perform some finite computation sequence constituting α. It has the opportunity to perform α such that φ results iff it has the opportunity to perform some finite computation sequence of α in such a way that φ results.

Even though the actions from $\mathrm{Ac}^{\oplus}$ are possibly nondeterministic, the finite computation sequences that constitute such an action are still deterministic.

4.7. Proposition. *All actions from* $\mathrm{Ac}^{\oplus}_{\mathrm{b}}$ *are deterministic in* $\mathbf{F}^{\oplus}$.

The nondeterministic action constructor $\oplus$ indeed behaves as desired, both for opportunity and result, as for ability.

4.8. Proposition. *For all* $i \in \mathrm{A}$, $\alpha_1, \alpha_2 \in \mathrm{Ac}^{\oplus}$ *and* $\varphi \in \mathrm{L}^{\oplus}$ *we have:*

- $\models^{\oplus} \langle \mathrm{do}_i(\alpha_1 \oplus \alpha_2)\rangle\varphi \leftrightarrow (\langle \mathrm{do}_i(\alpha_1)\rangle\varphi \vee \langle \mathrm{do}_i(\alpha_2)\rangle\varphi)$
- $\models^{\oplus} \mathbf{A}_i(\alpha_1 \oplus \alpha_2) \leftrightarrow (\mathbf{A}_i\alpha_1 \vee \mathbf{A}_i\alpha_2)$

The validities that were found in Proposition 3.5 to characterise the compositional behaviour of actions with respect to opportunities and results do also hold for $\models^{\oplus}$. Furthermore, the operator $[\mathrm{do}_i(\alpha)]$, with $i \in \mathrm{A}$ and $\alpha \in \mathrm{Ac}^{\oplus}$ is still normal.

4.9. Proposition. *For all* $i \in \mathrm{A}$, $\alpha, \alpha_1, \alpha_2 \in \mathrm{Ac}^{\oplus}$ *and* $\varphi, \psi \in \mathrm{L}^{\oplus}$ *we have:*

1. $\models^{\oplus} \langle \mathrm{do}_i(\mathtt{confirm}\,\varphi)\rangle\psi \leftrightarrow (\varphi \wedge \psi)$
2. $\models^{\oplus} \langle \mathrm{do}_i(\alpha_1;\alpha_2)\rangle\psi \leftrightarrow \langle \mathrm{do}_i(\alpha_1)\rangle\langle \mathrm{do}_i(\alpha_2)\rangle\psi$
3. $\models^{\oplus} \langle \mathrm{do}_i(\mathtt{if}\,\varphi\,\mathtt{then}\,\alpha_1\,\mathtt{else}\,\alpha_2\,\mathtt{fi})\rangle\psi \leftrightarrow ((\varphi \wedge \langle \mathrm{do}_i(\alpha_1)\rangle\psi) \vee (\neg\varphi \wedge \langle \mathrm{do}_i(\alpha_2)\rangle\psi))$
4. $\models^{\oplus} \langle \mathrm{do}_i(\mathtt{while}\,\varphi\,\mathtt{do}\,\alpha\,\mathtt{od})\rangle\psi \leftrightarrow ((\neg\varphi \wedge \psi) \vee (\varphi \wedge \langle \mathrm{do}_i(\alpha)\rangle\langle \mathrm{do}_i(\mathtt{while}\,\varphi\,\mathtt{do}\,\alpha\,\mathtt{od})\rangle\psi))$
5. $\models^{\oplus} [\mathrm{do}_i(\alpha)](\varphi \to \psi) \to ([\mathrm{do}_i(\alpha)]\varphi \to [\mathrm{do}_i(\alpha)]\psi)$
6. $\models^{\oplus} \psi \Rightarrow \models^{\oplus} [\mathrm{do}_i(\alpha)]\psi$

With regard to abilities, the behaviour of the sequential composition and the repetitive composition with respect to $\models^{\oplus}$ differs from the behaviour with respect to $\models$; for the other action constructors the same validities as given in Proposition 3.6 are found to hold for $\models^{\oplus}$.

4.10. Proposition. *For* $i \in \mathrm{A}$, $\alpha_1, \alpha_2 \in \mathrm{Ac}^{\oplus}$ *and* $\varphi \in \mathrm{L}^{\oplus}$ *we have:*

- $\models^{\oplus} \mathbf{A}_i\texttt{confirm}\,\varphi \leftrightarrow \varphi$
- $\models^{\oplus} \mathbf{A}_i\texttt{if}\,\varphi\,\texttt{then}\,\alpha_1\,\texttt{else}\,\alpha_2\,\texttt{fi} \leftrightarrow ((\varphi \wedge \mathbf{A}_i\alpha_1) \vee (\neg\varphi \wedge \mathbf{A}_i\alpha_2))$

4.11. Proposition. *For $i \in \mathrm{A}$, $\alpha, \alpha_1, \alpha_2, \alpha_3 \in \mathrm{Ac}^{\oplus}$ and $\varphi \in \mathrm{L}^{\oplus}$ we have:*

1. $\mathbf{A}_i\alpha_1;\alpha_2 \rightarrow [\mathrm{do}_i(\alpha_1)]\mathbf{A}_i\alpha_2$ *is not for all $\alpha_1, \alpha_2 \in \mathrm{Ac}^{\oplus}$, $\models^{\oplus}$-valid*
2. $\models^{\oplus} \mathbf{A}_i\alpha_1 \wedge [\mathrm{do}_i(\alpha_1)]\mathbf{A}_i\alpha_2 \rightarrow \mathbf{A}_i\alpha_1;\alpha_2$
3. $\models^{\oplus} \mathbf{A}_i(\alpha_1 \oplus \alpha_2);\alpha_3 \leftrightarrow \mathbf{A}_i(\alpha_1;\alpha_3) \vee \mathbf{A}_i(\alpha_2;\alpha_3)$
4. $\models^{\oplus} \mathbf{A}_i\alpha_1;(\alpha_2 \oplus \alpha_3) \leftrightarrow \mathbf{A}_i(\alpha_1;\alpha_2) \vee \mathbf{A}_i(\alpha_1;\alpha_3)$
5. $\models^{\oplus} \mathbf{A}_i\texttt{while}\,\varphi\,\texttt{do}\,\alpha\,\texttt{od} \leftrightarrow \neg\varphi \vee (\varphi \wedge \mathbf{A}_i\alpha;\texttt{while}\,\varphi\,\texttt{do}\,\alpha\,\texttt{od})$
6. $\mathbf{A}_i\texttt{while}\,\varphi\,\texttt{do}\,\alpha\,\texttt{od} \rightarrow (\neg\varphi \vee [\mathrm{do}_i(\alpha)]\mathbf{A}_i\texttt{while}\,\varphi\,\texttt{do}\,\alpha\,\texttt{od})$ *is not for all $\varphi \in \mathrm{L}^{\oplus}, \alpha \in \mathrm{Ac}^{\oplus}$, $\models^{\oplus}$-valid*
7. $\models^{\oplus} (\neg\varphi \vee (\varphi \wedge \mathbf{A}_i\alpha \wedge [\mathrm{do}_i(\alpha)]\mathbf{A}_i\texttt{while}\,\varphi\,\texttt{do}\,\alpha\,\texttt{od})) \rightarrow \mathbf{A}_i\texttt{while}\,\varphi\,\texttt{do}\,\alpha\,\texttt{od}$

The first two items of Proposition 4.11 state that, though truth of $\mathbf{A}_i\alpha_1 \wedge [\mathrm{do}_i(\alpha_1)]\mathbf{A}_i\alpha_2$ is still sufficient to conclude that $\mathbf{A}_i\alpha_1;\alpha_2$ is true, it is no longer necessary. This is obviously one of the desiderata as they follow from the observations made in Example 4.2. The same can be said for the items 3 and 4: the formal unravelling of actions indeed results in the intuitively desirable equivalences. Item 5 concerns an equivalence that is also valid for $\models$, and that may be seen as a slightly weakened variant of the formulae considered in the items 6 and 7. The invalidity formalised in item 6 is related to the invalidity given in item 1, and is in fact a direct consequence of the combination of item 1 and item 5.

4.2.1 Practical possibility and free choice

In the previous chapter we defined the Can-predicate and the Cannot-predicate to model the knowledge of agents on their practical (im)possibilities. The intuitive idea underlying the definition of these predicates is that practical possibility is stated in terms of the correctness and feasibility to bring about some proposition by performing some action. In Chapter 3 we propose to formalise correctness through the formula $\langle \mathrm{do}_i(\alpha)\rangle\varphi$, which expresses that α is correct for agent i to bring about φ. Feasibility is formalised through the formula $\mathbf{A}_i\alpha$, which states that α is feasible for agent i, i.e. within its capabilities. The notion of practical possibility is then formalised by $\mathbf{PracPoss}_i(\alpha,\varphi) \triangleq \mathbf{Correct}_i(\alpha,\varphi) \wedge \mathbf{Feasible}_i\alpha$. The Can-predicate and the Cannot-predicate are formalised by $\mathbf{Can}_i(\alpha,\varphi) \triangleq \mathbf{K}_i\mathbf{PracPoss}_i(\alpha,\varphi)$ and $\mathbf{Cannot}_i(\alpha,\varphi) \triangleq \mathbf{K}_i\neg\mathbf{PracPoss}_i(\alpha,\varphi)$, respectively. Although these definitions are intuitively perfectly acceptable when only deterministic actions are considered, they are no longer so when nondeterminism is involved.

4.12. Example. Suppose that the author of this thesis knows that he is given the opportunity to reach the top of K2 since somebody placed him at the foot of this mountain.

Although he knows that he has the opportunity, he also knows that he is not able to take it, because it is certainly not within his capabilities to climb K2. This situation may be formalised through the formula

$\mathbf{K}_b(\langle \mathrm{do}_b(\texttt{climb\_K2})\rangle \texttt{top\_reached} \wedge \neg \mathbf{A}_b \texttt{climb\_K2})$

Since the author knows for sure that he is specialised in doing nothing, the formula $\mathbf{K}_b\mathbf{A}_b\texttt{skip}$ is also definitely true. But then it follows that both

$\mathbf{K}_b\langle \mathrm{do}_b(\texttt{skip} \oplus \texttt{climb\_K2})\rangle \texttt{top\_reached}$

and

$\mathbf{K}_b\mathbf{A}_b(\texttt{skip} \oplus \texttt{climb\_K2})$

hold, which implies that $\mathbf{Can}_b(\texttt{skip} \oplus \texttt{climb\_K2}, \texttt{top\_reached})$ holds, i.e. the author of this thesis knows that he somehow has the practical possibility to reach the top of K2!

The situation observed in the example above resembles the one observed in Example 4.2, and could be taken as another sign that straightforward approaches towards nondeterminism do not work when both opportunities and abilities are involved. From Example 4.12 we learn the same lesson as from Example 4.2, namely that nondeterministic actions are to be unravelled, also when defining known practical (im)possibilities. One has to ensure that whenever the formula $\mathbf{Can}_i(\alpha, \varphi)$ holds, some finite computation sequence α' of α exists which is such that i has the reliable opportunity to perform α' such that φ results, while it also has the reliable ability to do α'. We bring this about by defining the Can-predicate and the Cannot-predicate as independent, primitive entities with their own semantics. In this semantics, which is based on the unravelling of actions, it is used that $\langle \mathrm{do}_i(\alpha)\rangle\varphi \wedge \mathbf{A}_i\alpha$ does still represent practical possibility for finite computation sequences, which after all are deterministic actions.

4.13. Definition. The binary relation $\models^{\oplus}$ between a formula from $\mathrm{L}^{\oplus}$ and a pair M, s consisting of a model M for $\mathrm{L}^{\oplus}$ and a state s in M is for the Can-predicate and the Cannot-predicate defined by:

$$\begin{array}{ll} \mathrm{M}, s \models^{\oplus} \mathbf{Can}_i(\alpha, \varphi) & \Leftrightarrow \forall s' \in [s]_{\mathrm{R}(i)} \exists \alpha' \in \mathrm{CS}^{\oplus}(\alpha)(\mathrm{M}, s' \models^{\oplus} \mathbf{PracPoss}_i(\alpha', \varphi)) \\ \mathrm{M}, s \models^{\oplus} \mathbf{Cannot}_i(\alpha, \varphi) & \Leftrightarrow \forall s' \in [s]_{\mathrm{R}(i)} \forall \alpha' \in \mathrm{CS}^{\oplus}(\alpha)(\mathrm{M}, s' \models^{\oplus} \neg\mathbf{PracPoss}_i(\alpha', \varphi)) \end{array}$$

where $\mathbf{PracPoss}_i(\alpha, \varphi)$ is defined as in Chapter 3, i.e. $\mathbf{PracPoss}_i(\alpha, \varphi) \triangleq \langle \mathrm{do}_i(\alpha)\rangle\varphi \wedge \mathbf{A}_i\alpha$.

Interpreting the Can-predicate and the Cannot-predicate as in Definition 4.13 indeed solves the counterintuitive situation encountered in Example 4.12. For not only does $\mathbf{Can}_b(\texttt{skip} \oplus \texttt{climb\_K2}, \texttt{top\_reached})$ no longer hold, it is even the case that the stronger formula $\mathbf{Cannot}_b(\texttt{skip} \oplus \texttt{climb\_K2}, \texttt{top\_reached})$ holds, i.e. the author of this thesis knows that he does not have the practical possibility to reach the top of K2 by choosing

freely between doing nothing and climbing the mountain. More in general, the Can-predicate and the Cannot-predicate indeed show an intuitively acceptable behaviour when ranging over nondeterministic actions.

4.14. Proposition. *For* $i \in \mathrm{A}, \alpha_1, \alpha_2 \in \mathrm{Ac}^{\oplus}$ *and* $\varphi \in \mathrm{L}^{\oplus}$ *we have:*

- $\models^{\oplus} \mathbf{Can}_i(\alpha_1 \oplus \alpha_2, \varphi) \leftrightarrow \mathbf{Can}_i(\alpha_1, \varphi) \vee \mathbf{Can}_i(\alpha_2, \varphi)$
- $\models^{\oplus} \mathbf{Cannot}_i(\alpha_1 \oplus \alpha_2, \varphi) \leftrightarrow \mathbf{Cannot}_i(\alpha_1, \varphi) \wedge \mathbf{Cannot}_i(\alpha_2, \varphi)$

Proposition 4.14 clearly expresses the idea that the agent knows that it is in control when performing an internal nondeterministic action: the agent knows that an internal nondeterministic choice between two actions is correct and feasible if and only if one of the actions to be chosen is. Since the choice is completely free to the agent, this is what one intuitively would expect.

Since the Can-predicate and the Cannot-predicate as formalised in Definition 4.13 are essentially different from the predicates formalised in Section 3.3, it is both interesting and important to look at the differences and similarities between the two kinds of definitions. To this end we reconsider the validities found in Section 3.3 to characterise the Can-predicate and the Cannot-predicate. In general one can say that the validities concerning confirmations or conditional compositions as found in Proposition 3.18 do also hold for $\models^{\oplus}$ when defining the Can-predicate and the Cannot-predicate as above. For the sequential composition and the repetitive composition we have the following.

4.15. Proposition. *The following formulae are* <u>*not*</u> *for all* $i \in \mathrm{A}$, $\alpha_1, \alpha_2 \in \mathrm{Ac}^{\oplus}$ *and* $\varphi, \psi \in \mathrm{L}^{\oplus}$ $\models^{\oplus}$*-valid:*

1. $\mathbf{Can}_i(\alpha_1, \langle \mathrm{do}_i(\alpha_2) \rangle \varphi \wedge \mathbf{A}_i \alpha_2) \rightarrow \mathbf{Can}_i(\alpha_1; \alpha_2, \varphi)$
2. $\mathbf{Cannot}_i(\alpha_1; \alpha_2, \varphi) \rightarrow \mathbf{Cannot}_i(\alpha_1, \langle \mathrm{do}_i(\alpha_2) \rangle \varphi \wedge \mathbf{A}_i \alpha_2)$
3. $\mathbf{Can}_i(\alpha, \langle \mathrm{do}_i(\texttt{while}\, \varphi\, \texttt{do}\, \alpha\, \texttt{od}) \rangle \psi \wedge \mathbf{A}_i \texttt{while}\, \varphi\, \texttt{do}\, \alpha\, \texttt{od}) \wedge \mathbf{K}_i \varphi \rightarrow$
   $\mathbf{Can}_i(\texttt{while}\, \varphi\, \texttt{do}\, \alpha\, \texttt{od}, \psi) \wedge \mathbf{K}_i \varphi$

The non-validities stated in Proposition 4.15 essentially follow from the fact that for nondeterministic actions the combination of $\langle \mathrm{do}_i(\alpha) \rangle \varphi \wedge \mathbf{A}_i \alpha$ does not represent practical possibility. That is, even though it states that the agent has the reliable opportunity to perform α in such a way that φ results while it also has the reliable ability to perform α, it fails to tie this opportunity and ability together: the agent may have the reliable opportunity to do α in some way while not having the ability to do α in the same way (but in another way). A similar observation underlies Example 4.12.

With respect to the end-part of the Can-predicate and the Cannot-predicate it turns out that most of the formulae given in Proposition 3.19 are also valid for $\models^{\oplus}$. The most important exceptions are the following ones.

4.16. PROPOSITION. *For $i \in \mathrm{A}$ we have:*

- $\mathbf{Can}_i(\alpha, \varphi) \wedge \mathbf{Can}_i(\alpha, \neg\varphi)$ *is* $\models^{\oplus}$*-satisfiable for certain* $\alpha \in \mathrm{Ac}^{\oplus}$, $\varphi \in \mathrm{L}^{\oplus}$
- $\mathbf{Can}_i(\alpha, \varphi) \wedge \mathbf{Can}_i(\alpha, \psi) \rightarrow \mathbf{Can}_i(\alpha, \varphi \wedge \psi)$ *is not for all* $\alpha \in \mathrm{Ac}^{\oplus}$, $\psi \in \mathrm{L}^{\oplus}$, $\models^{\oplus}$*-valid*

The items of Proposition 4.16 nicely indicate both the nondeterministic character of the actions from $\mathrm{Ac}^{\oplus}$ and the fact that the agent itself is in control. The first item states that it is possible that an agent knows that it has both the practical possibility to perform α in such a way that φ results as well as the practical possibility to perform α in such a way that $\neg\varphi$ results. Since the agent itself decides *how* an internal nondeterministic action is performed, satisfiability of this formula seems to make sense. Note that it is a *conditio sine qua non* that α is nondeterministic for $\mathbf{Can}_i(\alpha, \varphi) \wedge \mathbf{Can}_i(\alpha, \neg\varphi)$ to be satisfiable. The invalidity of the second item follows directly from the satisfiability of the first item: even though $\mathbf{Can}_i(\alpha, \varphi) \wedge \mathbf{Can}_i(\alpha, \neg\varphi)$ may hold, $\mathbf{Can}_i(\alpha, \varphi \wedge \neg\varphi)$ will never be true.

4.3 External nondeterminism

In this section we propose two possible approaches towards a formalisation of external nondeterministic actions. In the first of these it is assumed that the agents are optimistic in the sense of the previous chapter, i.e. agents are omnipotent in the counterfactual state of affairs. This formalisation of external nondeterminism for optimistic agents, though slightly more complex, resembles the one given for internal nondeterminism in the previous section. In particular, this semantics is also based on the unravelling of actions, though in a more elaborate and subtle way. When introducing external nondeterminism for pessimistic agents, i.e. for agents that are assumed to be nilpotent in the counterfactual state of affairs, it is not necessary to unravel actions in order to come up with an adequate semantics. Instead we propose a semantics that is inspired by, and uses elements of, the one presented by Peleg [101] for Concurrent Propositional Dynamic Logic (CPDL).

The external nondeterministic combination of two actions α_1 and α_2 is denoted by $\alpha_1 + \alpha_2$. Informally, execution of $\alpha_1 + \alpha_2$ corresponds to execution of either α_1 or of α_2, where the external environment decides which action to perform.

4.17. DEFINITION. To define the language L^{+}, the alphabet is extended with the action constructor $\_ + \_$, representing the external nondeterministic choice operator. The language L^{+} is the smallest superset of Π closed under the core clauses. The class Ac^{+} of actions is the smallest superset of At closed under the core clauses and such that $\alpha_1 + \alpha_2 \in \mathrm{Ac}^{+}$ whenever $\alpha_1 \in \mathrm{Ac}^{+}, \alpha_2 \in \mathrm{Ac}^{+}$.

Just as the language $\mathrm{L}^{\oplus}$ defined in the previous section, L^{+} is easily seen to be a genuine extension of the language L defined in the previous chapter.

To treat actions containing the constructor $+$ as fully-fledged ones, we have to decide on how to interpret and formalise the ability, opportunity and result for external nondeterministic actions, given the intuitive idea that the external environment, which makes the choice as to what action to perform, may display a demonic behaviour. Starting from this principle, it seems reasonable to declare an agent to have the opportunity to perform the action $\alpha_1 + \alpha_2$ in such a way that φ is certain to result iff it has both the reliable opportunity to do α_1 such that φ results and it has the reliable opportunity to perform α_2 in such a way that execution leads to φ. For the agent has to expect the worst from the external environment, and therefore should be prepared both to having to execute α_1 and to having to execute α_2. Formally this amounts to the formula

$$\langle \mathrm{do}_i(\alpha_1 + \alpha_2)\rangle\varphi \leftrightarrow \langle \mathrm{do}_i(\alpha_1)\rangle\varphi \wedge \langle \mathrm{do}_i(\alpha_2)\rangle\varphi \qquad (\Diamond)$$

being valid. A direct consequence of this formula being valid is that

$$[\mathrm{do}_i(\alpha_1 + \alpha_2)]\varphi \leftrightarrow [\mathrm{do}_i(\alpha_1)]\varphi \vee [\mathrm{do}_i(\alpha_2)]\varphi \qquad (\Box)$$

is also valid. By a similar argument one comes to the conclusion that an agent is reliably capable of performing the action $\alpha_1 + \alpha_2$ iff it is both reliably capable of performing α_1 and it is reliably capable of performing α_2. From a formal point of view this leads to the equivalence

$$\mathbf{A}_i(\alpha_1 + \alpha_2) \leftrightarrow \mathbf{A}_i\alpha_1 \wedge \mathbf{A}_i\alpha_2 \qquad (\blacklozenge)$$

being valid. Just as we did for internal nondeterminism, we investigate the consequences of combining these equivalences with the ones describing the optimistic and the pessimistic approach formalised in the previous chapter. It turns out that in the optimistic approach counterintuitive situations arise. Whereas one intuitively expects having the ability to perform an action $(\alpha_1 + \alpha_2); \alpha_3$ to be equivalent to having both the ability to do $\alpha_1; \alpha_3$ and having the ability to do $\alpha_2; \alpha_3$, this equivalence does not come about when combining $(\Box)$ and $(\blacklozenge)$ given above according to the optimistic approach of Chapter 3.

4.18. Example. Assume an agent i to be in a state such that $\mathbf{A}_i a_1, \mathbf{A}_i a_2, \langle \mathrm{do}_i(a_1)\rangle\neg\mathbf{A}_i a_3$ and $[\mathrm{do}_i(a_2)]\bot$ are true in the state, for $a_1, a_2, a_3 \in \mathrm{At}$. In the optimistic approach of the previous chapter, it is concluded that the agent, though able to do $a_2; a_3$, is not able to perform $a_1; a_3$, which would intuitively suffice to conclude that it is not able to perform $(a_1 + a_2); a_3$. However, using $(\blacklozenge)$ one concludes from $\mathbf{A}_i a_1$ and $\mathbf{A}_i a_2$ that also $\mathbf{A}_i(a_1 + a_2)$ holds. Furthermore, from $[\mathrm{do}_i(a_2)]\bot$ it follows by $(\Box)$ that $[\mathrm{do}_i(a_1 + a_2)]\bot$ holds and hence that $[\mathrm{do}_i(a_1 + a_2)]\mathbf{A}_i a_3$ holds. Combining $\mathbf{A}_i(a_1 + a_2)$ with $[\mathrm{do}_i(a_1 + a_2)]\mathbf{A}_i a_3$ forces one to conclude that $\mathbf{A}_i(a_1 + a_2); a_3$ holds. That is, even though the agent is not able to do $a_1; a_3$, it is still concluded that it is able to do $(a_1 + a_2); a_3$.

To deal with the problems observed in Example 4.18, we propose an approach similar to the one used to solve the problems that occur with the internal nondeterministic choice, i.e. an approach based on the unravelling of actions.

It turns out that straightforwardly adding the equivalences $(\Diamond)$ and $(\blacklozenge)$ to the ones describing the pessimistic approach of Chapter 3 does not lead to counterintuitive situations. For instance, assuming the equivalences $(\Diamond)$ and $(\blacklozenge)$ to be valid, it can be shown in the pessimistic approach that $\mathbf{A}_i(\alpha_1 + \alpha_2); \alpha_3$ is indeed equivalent with $\mathbf{A}_i(\alpha_1; \alpha_3) \wedge \mathbf{A}_i(\alpha_2; \alpha_3)$. Thus, defining a semantics that respects the validities of Chapter 3 for the pessimistic approach in combination with $(\Diamond)$ and $(\blacklozenge)$ constitutes the main part of our formalisation of the external nondeterministic choice for pessimistic agents.

4.3.1 External nondeterminism and optimistic agents

The semantics used to interpret external nondeterministic actions for optimistic agents is also based on the idea of unravelling an action into the finite-length sequences of semi-atomic actions that constitute it. Although the unravelling of external nondeterministic actions is a little more involved than that of internal nondeterministic actions, it is still the case that also the former kind of unravelling is based on the unravel function introduced in Chapter 2.

4.19. Definition. The function $\mathrm{CS}^+ : \mathrm{Ac}^+ \to \wp(\mathrm{Ac}_\mathrm{b}^+)$ is for the non-core action constructor $+$ defined by: $\mathrm{CS}^+(\alpha_1 + \alpha_2) = \mathrm{CS}^+(\alpha_1) \cup \mathrm{CS}^+(\alpha_2)$.

The models for L^+, just as those for $\mathrm{L}^\oplus$, need not interpret other primitive operators besides the core ones, and therefore contain nothing but the core elements.

4.20. Definition. A model M for the language L^+ is a tuple consisting of the core elements. The class of all models for L^+ is denoted by $\mathbf{M}^+$.

Inspired by the formalisation of internal nondeterministic actions as presented in Section 4.2, one could be tempted to interpret non-core formulae containing external nondeterministic actions in the following manner:

$$\mathrm{M}, s \models^+ \langle \mathrm{do}_i(\alpha) \rangle \varphi \Leftrightarrow \forall \alpha' \in \mathrm{CS}^+(\alpha) \exists s' \in \mathrm{S}(\mathtt{r}^+(i, \alpha')(s) = s' \,\&\, \mathrm{M}, s' \models^+ \varphi) \qquad (\star)$$

$$\mathrm{M}, s \models^+ \mathbf{A}_i \alpha \quad \Leftrightarrow \forall \alpha' \in \mathrm{CS}^+(\alpha)(\mathtt{c}^+(i, \alpha')(s) = \mathbf{1}) \qquad (\star\star)$$

where $\mathtt{r}^+$ and $\mathtt{c}^+$ are defined as $\mathtt{r}^\oplus$ and $\mathtt{c}^\oplus$, respectively. However, as shown in the following example, adopting a semantics based on $(\star)$ and $(\star\star)$ may lead to counterintuitive situations.

4.21. EXAMPLE. Consider the action $\alpha \triangleq \texttt{if}\ \top\ \texttt{then skip else fail fi}$. Given the intuitive meaning of the conditional composition, one expects α to behave as the action skip. However, the equivalences $(\star)$ and $(\star\star)$ as given above are not adequate to ensure this. For $\texttt{confirm}\,\neg\top;\texttt{fail}$ is an element of $\mathrm{CS}^+(\alpha)$, and hence neither $\langle \mathrm{do}_i(\alpha)\rangle\top$ nor $\mathbf{A}_i\alpha$ is satisfiable according to $(\star)$ and $(\star\star)$.

Example 4.21 clearly shows the problems associated with using $(\star)$ and $(\star\star)$ to interpret the non-core formulae. The function CS^+ is in fact not sufficiently discriminating to define the finite sequences of semi-atomic actions that constitute the execution of a given action. Even though the sequences that do constitute the execution of an action α are elements of the set $\mathrm{CS}^+(\alpha)$[1], this set furthermore contains some sequences that will in certain circumstances not occur in the halting execution of α and therefore should be left out of consideration. Since these sequences do not constitute the execution of the action, it is highly unreasonable to demand the agent to have the opportunity or ability to perform these sequences in order for it to have the opportunity or ability to perform the action. For example, the only relevant computation sequence of the action $\texttt{if}\ \top\ \texttt{then skip else fail fi}$ considered in Example 4.21 is $\texttt{confirm}\,\top;\texttt{skip}$, since $\texttt{confirm}\,\neg\top;\texttt{fail}$ will never occur in an execution of $\texttt{if}\ \top\ \texttt{then skip else fail fi}$. Hence it should be both necessary and sufficient that an agent has the opportunity or ability to do $\texttt{confirm}\,\top;\texttt{skip}$ in order to conclude that it has the opportunity or ability to perform $\texttt{if}\ \top\ \texttt{then skip else fail fi}$. Analogously, in states where some formula φ holds, the relevant computation sequencesof an action $\texttt{if}\ \varphi\ \texttt{then}\ \alpha_1\ \texttt{else}\ \alpha_2\ \texttt{fi}$ are contained in those of α_1; the finite computation sequences of α_2 are not relevant in these states.

To single out the finite computation sequences that are *relevant* given an agent and a state of a model, we introduce the notion of *finite computation runs*. The finite computation runs of an action α, for an agent i in a state s, are exactly those finite computation sequences of α that jointly constitute α for i in s, i.e. i has the opportunity or ability to do α iff it has the opportunity or ability to do any of the finite computation runs of α.

As we already mentioned in Chapter 3, no agent is assumed to have the ability to perform an action for which execution does not terminate, i.e. executing the action would take infinite time. There is another source of non-termination, in which execution of an action does not even begin and therefore does also not terminate. For the purposes of this chapter, we refer to the first kind of non-termination as *infinite non-termination*. A typical example of an infinitely non-terminating event is $\mathrm{do}_i(\texttt{while}\ \top\ \texttt{do skip od})$. The second kind of non-termination is called *void non-termination*; a typical example of a

[1] This is exactly the reason why the use of $\mathrm{CS}^{\oplus}$ in Definition 4.5 does not give any problems. For the relevant sequences which are indeed present in $\mathrm{CS}^{\oplus}(\alpha)$ are the ones 'found' by the existential quantification over $\mathrm{CS}^{\oplus}(\alpha)$.

voidly non-terminating event is $\mathrm{do}_i(\mathtt{fail})$. Note that it is possible that optimistic agents are able to perform actions that would lead to voidly non-terminating events; hence it is possible that actions resulting in a state s for an agent i in voidly non-terminating events still are relevant. In defining the function computing the finite computation runs of actions we therefore have to discriminate between infinitely and voidly non-terminating events. With respect to the former, the definition of finite computation runs is such that if some computation sequence of an action α results, for a given agent i in a given state s, in an infinitely non-terminating event, the set of finite computation runs of α for the agent and the state is defined to be equal to the singleton set $\{\mathtt{fail}\}$. If none of the finite computation sequences of action α results in an infinitely non-terminating event for i in s, the set of finite computation runs of α is defined inductively. In this inductive definition it is taken into account that, depending on the truth or falsity of the condition, only some of the finite computation sequences of a conditional or a repetitive composition are finite computation runs. Note in particular that the set of finite computation runs of an action is a situated notion, dependent on the agent executing the action and the state in which it is executed, which is, in contrast with the set of finite computation sequences, no longer determined by syntax alone. The set of finite computation runs of α for i in s is denoted by $\mathrm{CR}^{+}_{\mathrm{M}}(i, \alpha, s)$.

To check whether an action α, for a given agent i and a state s, can be executed in such a way that an infinitely non-terminating event results, the termination predicate $\mathrm{Ter}_{\mathrm{M}}$ is used. The definition of this predicate is a rather straightforward formalisation of the idea that infinite events result in infinitely many state-transitions. If we assume that execution of semi-atomic actions takes one execution cycle, then execution of an action that results in an infinitely non-terminating event takes more than k execution cycles, for all $k \in \mathbb{N}$. To be able to deal with infinite while-loops, we incorporate the prefix relation in the definition of the $\mathrm{Ter}_{\mathrm{M}}$ predicate. If we consider for instance the action $\alpha \triangleq \mathtt{while}\ \top\ \mathtt{do}\ \mathtt{skip}\ \mathtt{od}$ in a given state s for a given agent i, it is obvious that execution of α by i results in infinitely many state-transitions from s to s. However, the set of finite computation sequences of α fails to express this: since all finite computation sequences of α are of the form $\alpha'; \mathtt{confirm}\ \bot$, all of these computation sequences result in voidly non-terminating events! The definition of $\mathrm{Ter}_{\mathrm{M}}$ as we give it, i.e. using the prefix relation, can cope with the situation sketched above: it expresses that for all natural numbers k a natural number $l \geq k$ exists such that for some finite computation sequence $\alpha' \in \mathrm{CS}(\mathtt{while}\ \top\ \mathtt{do}\ \mathtt{skip}\ \mathtt{od})$ the prefix of α' of length l results in l state-transitions from s to s, and hence the event $\mathrm{do}_i(\mathtt{while}\ \top\ \mathtt{do}\ \mathtt{skip}\ \mathtt{od})$ is infinitely non-terminating in all states s, for all agents i.

After this, hopefully explanatory, introduction the actual definitions of the various predicates can be given.

4.22. DEFINITION. The binary relation $\models^+$ between a formula from L^+ and a pair M, s consisting of a model M for L^+ and a state s in M is for the non-core formulae defined as follows:

$$\mathrm{M}, s \models^+ \langle \mathrm{do}_i(\alpha) \rangle \varphi \Leftrightarrow \forall \alpha' \in \mathrm{CR}^+_{\mathrm{M}}(i, \alpha, s) \exists s' \in \mathrm{S}(\mathtt{r}^+(i, \alpha')(s) = s' \,\&\, \mathrm{M}, s' \models^+ \varphi)$$
$$\mathrm{M}, s \models^+ \mathbf{A}_i \alpha \quad \Leftrightarrow \forall \alpha' \in \mathrm{CR}^+_{\mathrm{M}}(i, \alpha, s)(\mathtt{c}^+(i, \alpha')(s) = \mathbf{1})$$

where the functions $\mathrm{CR}^+_{\mathrm{M}}$, $\mathrm{Ter}_{\mathrm{M}}$, $\mathtt{r}^+$ and $\mathtt{c}^+$ are defined as follows.

$$\begin{array}{lcl}
\mathrm{CR}^+_{\mathrm{M}} & : & \mathrm{A} \times \mathrm{Ac}^+ \times \mathrm{S}^\cdot \to \wp(\mathrm{Ac}^+_{\mathrm{b}}) \\
\mathrm{CR}^+_{\mathrm{M}}(i, \alpha, s) & = & \{\mathtt{fail}\} \text{ if } \mathrm{Ter}_{\mathrm{M}}(i, \alpha, s) = \mathbf{0} \\
\multicolumn{3}{l}{\text{else if } \mathrm{Ter}_{\mathrm{M}}(i, \alpha, s) = \mathbf{1}\text{:}} \\
\mathrm{CR}^+_{\mathrm{M}}(i, \alpha, s) & = & \{\alpha\} \text{ if } \alpha \in \mathrm{Ac}^+_{\mathrm{s}} \\
\mathrm{CR}^+_{\mathrm{M}}(i, \alpha_1; \alpha_2, s) & = & \{\alpha'_1; \alpha'_2 \mid \alpha'_1 \in \mathrm{CR}^+_{\mathrm{M}}(i, \alpha_1, s), \\
 & & \qquad \alpha'_2 \in \mathrm{CR}^+_{\mathrm{M}}(i, \alpha_2, \mathtt{r}^+(i, \alpha'_1)(s))\} \\
\mathrm{CR}^+_{\mathrm{M}}(i, \mathtt{if}\, \varphi\, \mathtt{then}\, \alpha_1\, \mathtt{else}\, \alpha_2\, \mathtt{fi}, s) & = & \mathrm{CR}^+_{\mathrm{M}}(i, \mathtt{confirm}\, \varphi; \alpha_1, s) \text{ if } \mathrm{M}, s \models^+ \varphi \\
 & = & \mathrm{CR}^+_{\mathrm{M}}(i, \mathtt{confirm}\, \neg\varphi; \alpha_2, s) \text{ otherwise} \\
\mathrm{CR}^+_{\mathrm{M}}(i, \mathtt{while}\, \varphi\, \mathtt{do}\, \alpha\, \mathtt{od}, s) & = & \{\mathtt{confirm}\, \neg\varphi\} \text{ if } \mathrm{M}, s \not\models^+ \varphi \\
 & = & \{\alpha' \in \mathrm{CS}^+(\mathtt{while}\, \varphi\, \mathtt{do}\, \alpha\, \mathtt{od}) \mid \\
 & & \qquad (\alpha' = (\mathtt{confirm}\, \varphi; \beta_1); \beta),\ \beta_1 \in \mathrm{CR}^+_{\mathrm{M}}(i, \alpha, s), \\
 & & \qquad \beta \in \mathrm{CR}^+_{\mathrm{M}}(i, \mathtt{while}\, \varphi\, \mathtt{do}\, \alpha\, \mathtt{od}, \mathtt{r}^+(i, \beta_1)(s))\} \\
 & & \qquad \text{if } \mathrm{M}, s \models^+ \varphi \\
\mathrm{CR}^+_{\mathrm{M}}(i, \alpha_1 + \alpha_2, s) & = & \mathrm{CR}^+_{\mathrm{M}}(i, \alpha_1, s) \cup \mathrm{CR}^+_{\mathrm{M}}(i, \alpha_2, s) \\
\mathrm{CR}^+_{\mathrm{M}}(i, \alpha, \emptyset) & = & \mathrm{CS}^+(\alpha)
\end{array}$$

$$\begin{array}{lcl}
\mathrm{Ter}_{\mathrm{M}} & : & \mathrm{A} \times \mathrm{Ac}^+ \times \mathrm{S} \to \mathsf{bool} \\
\mathrm{Ter}_{\mathrm{M}}(i, \alpha, s) = \mathbf{0} & \Leftrightarrow & \forall k \in \mathbb{N} \exists l \in \mathbb{N} \exists \alpha_1 \in \mathrm{CS}^+(\alpha) \exists \alpha_2 \in \mathrm{Ac}^+_{\mathrm{b}}(l \geq k \,\&\, \alpha_2 = |\alpha_1|^+_l \,\& \\
 & & \qquad \mathtt{r}^+(i, \alpha_2)(s) \neq \emptyset)
\end{array}$$

$$\begin{array}{lcl}
\mathtt{r}^+ & : & \mathrm{A} \times \mathrm{Ac}^+_{\mathrm{b}} \to \mathrm{S}^\cdot \to \mathrm{S}^\cdot \\
\mathtt{r}^+(i, a)(s) & = & \mathtt{r}_0(i, a)(s) \\
\mathtt{r}^+(i, \mathtt{confirm}\, \varphi)(s) & = & \{s\} \text{ if } \mathrm{M}, s \models^+ \varphi \\
 & = & \emptyset \text{ otherwise} \\
\mathtt{r}^+(i, \alpha_1; \alpha_2)(s) & = & \mathtt{r}^+(i, \alpha_2)(\mathtt{r}^+(i, \alpha_1)(s)) \\
\mathtt{r}^+(i, \alpha)(\emptyset) & = & \emptyset
\end{array}$$

$$\begin{array}{lcl}
\mathtt{c}^+ & : & \mathrm{A} \times \mathrm{Ac}^+_{\mathrm{b}} \to \mathrm{S}^\cdot \to \mathsf{bool} \\
\mathtt{c}^+(i, a)(s) & = & \mathtt{c}_0(i, a)(s) \\
\mathtt{c}^+(i, \mathtt{confirm}\, \varphi)(s) & = & \mathbf{1} \text{ iff } \mathrm{M}, s \models^+ \varphi \\
\mathtt{c}^+(i, \alpha_1; \alpha_2)(s) & = & \mathbf{1} \text{ iff } \mathtt{c}^+(i, \alpha_1)(s) = \mathbf{1} \,\&\, \mathtt{c}^+(i, \alpha_2)(\mathtt{r}^+(i, \alpha_1)(s)) = \mathbf{1} \\
\mathtt{c}^+(i, \alpha)(\emptyset) & = & \mathbf{1}
\end{array}$$

Note that the definition of $\mathrm{Ter}_{\mathrm{M}}$ as given above is correct only since the nondeterminism that is considered here is *bounded*, i.e. no infinite branching inside a nondeterministic choice is possible (cf. [2]). It is not hard to see that for bounded nondeterminism it indeed holds that $\mathrm{Ter}_{\mathrm{M}}(i,\alpha,s) = \mathbf{0}$ implies that $\mathrm{do}_i(\alpha)$ is infinitely non-terminating in s. For nondeterminism which is not bounded this implication is in general not valid[2]. Note furthermore that the case distinction on $\mathrm{Ter}_{\mathrm{M}}(i,\alpha,s)$ ensures well-definedness of $\mathrm{CR}^+_{\mathrm{M}}(i,\alpha,s)$ in Definition 4.22: if $\mathrm{Ter}_{\mathrm{M}}(i,\alpha,s) = \mathbf{0}$ then $\mathrm{CR}^+_{\mathrm{M}}(i,\alpha,s)$ equals the well-defined set $\{\mathtt{fail}\}$, and if $\mathrm{Ter}_{\mathrm{M}}(i,\alpha,s) = \mathbf{1}$ then the inductive definition of $\mathrm{CR}^+_{\mathrm{M}}(i,\alpha,s)$ is indeed correct. As a last remark, note that the definition of $\mathrm{CR}^+_{\mathrm{M}}(i,\alpha,\emptyset)$ is somewhat arbitrary: replacing $\mathrm{CS}^+(\alpha)$ by for instance $\mathrm{Ac}^+_{\mathrm{b}}$ or by $\{\mathtt{skip}\}$ would not affect the $\models^+$-relation. However, the definition as it is given is such that some interesting and intuitively desirable relations exist between the set of finite computation runs and the set of finite computation sequences. Some of these relations are given in the following proposition, which furthermore summarises some properties of the $\mathrm{Ter}_{\mathrm{M}}$ predicate.

4.23. Proposition. *For all* $\mathrm{M} \in \mathbf{M}^+$, s, s' *in* M, $i \in \mathrm{A}$, $\alpha, \alpha_1, \alpha_2 \in \mathrm{Ac}^+$ *and* $\varphi \in \mathrm{L}^+$*:*

1. $\mathrm{Ter}_{\mathrm{M}}(i,\alpha_1;\alpha_2,s) = \mathbf{1} \Rightarrow \mathrm{Ter}_{\mathrm{M}}(i,\alpha_1,s) = \mathbf{1}\,\&$
   $\forall s'(\exists \alpha_1' \in \mathrm{CS}^+(\alpha_1))(s' = \mathtt{r}^+(i,\alpha_1')(s)) \Rightarrow \mathrm{Ter}_{\mathrm{M}}(i,\alpha_2,s') = \mathbf{1})$
2. $\mathrm{Ter}_{\mathrm{M}}(i, \mathtt{if}\,\varphi\,\mathtt{then}\,\alpha_1\,\mathtt{else}\,\alpha_2\,\mathtt{fi}, s) = \mathbf{1} \Rightarrow$
   $\mathrm{Ter}_{\mathrm{M}}(i, \mathtt{confirm}\,\varphi;\alpha_1, s) = \mathbf{1}\,\&\,\mathrm{Ter}_{\mathrm{M}}(i, \mathtt{confirm}\,\neg\varphi;\alpha_2, s) = \mathbf{1}$
3. $\mathrm{Ter}_{\mathrm{M}}(i, \mathtt{while}\,\varphi\,\mathtt{do}\,\alpha\,\mathtt{od}, s) = \mathbf{1} \Rightarrow$
   $\forall \alpha'' \in \mathrm{Ac}^+_{\mathrm{b}}(\exists \alpha' \in \mathrm{CS}^+(\mathtt{while}\,\varphi\,\mathtt{do}\,\alpha\,\mathtt{od})(\mathrm{Prefix}^+(\alpha'',\alpha')) \Rightarrow$
   $(\mathtt{r}^+(i,\alpha'')(s) = s' \Rightarrow \mathrm{Ter}_{\mathrm{M}}(i, \mathtt{confirm}\,\varphi;\alpha, s') = \mathbf{1}))$
4. $\mathrm{Ter}_{\mathrm{M}}(i,\alpha_1+\alpha_2,s) = \mathbf{1} \Rightarrow \mathrm{Ter}_{\mathrm{M}}(i,\alpha_1,s) = \mathbf{1}\,\&\,\mathrm{Ter}_{\mathrm{M}}(i,\alpha_2,s) = \mathbf{1}$
5. $\mathrm{CS}^+(\alpha) \neq \emptyset$ *and* $\mathrm{CR}^+_{\mathrm{M}}(i,\alpha,s) \neq \emptyset$
6. $\mathrm{Ter}_{\mathrm{M}}(i,\alpha,s) = \mathbf{1} \Rightarrow \mathrm{CR}^+_{\mathrm{M}}(i,\alpha,s) \subseteq \mathrm{CS}^+(\alpha)$
7. $\mathrm{Ter}_{\mathrm{M}}(i,\alpha,s) = \mathbf{1} \Rightarrow \forall \alpha' \in \mathrm{Ac}^+_{\mathrm{b}} \forall s'(\alpha' \in \mathrm{CS}^+(\alpha)\,\&\,\mathtt{r}^+(i,\alpha')(s) = s' \Leftrightarrow$
   $\alpha' \in \mathrm{CR}^+_{\mathrm{M}}(i,\alpha,s)\,\&\,\mathtt{r}^+(i,\alpha')(s) = s')$
8. $\mathrm{Ter}_{\mathrm{M}}(i,\alpha,s) = \mathbf{1} \Rightarrow \forall \alpha' \in \mathrm{Ac}^+_{\mathrm{b}}(\alpha' \in \mathrm{CS}^+(\alpha)\,\&\,\mathtt{c}^+(i,\alpha')(s) = \mathbf{1} \Leftrightarrow$
   $\alpha' \in \mathrm{CR}^+_{\mathrm{M}}(i,\alpha,s)\,\&\,\mathtt{c}^+(i,\alpha')(s) = \mathbf{1})$

The first four items of Proposition 4.23 show that the termination predicate behaves as desired for composite actions. That is, a sequentially composed action terminates only if the first part of the sequence terminates and the second part terminates no matter how the first part has been performed, a conditional composition terminates only if confirming the condition followed by the then-part terminates and so does confirming

[2] As an example, assume that $\alpha_j \in \mathrm{Ac}^+_{\mathrm{b}}$ is for some model M with state s and agent i such that $|\alpha_j|^+ = j$ and $\mathtt{r}^+(i,\alpha_j)(s) \neq \emptyset$, for all $j \in \mathbb{N}$. Consider the countably nondeterministic action $\alpha = \alpha_1 + \alpha_2 + \alpha_3 + \ldots$. Even though the event $\mathrm{do}_i(\alpha)$ is not finitely non-terminating in s, it still holds that $\mathrm{Ter}_{\mathrm{M}}(i,\alpha,s) = \mathbf{0}$.

the negation of the condition followed by the else-part, and an external nondeterministic choice terminates only if both constituents of the choice terminate. Finally, a repetitive composition terminates only if execution of the body of the while-loop terminates in all points during execution of the loop. The last four items of Proposition 4.23 compare and relate the set of finite computation sequences to the set of finite computation runs of a given action. The fifth item states that both sets are nonempty, for each action α. With respect to the set of finite computation runs this result could be interpreted as stating that for every action at least one relevant finite computation sequence exists, which implies that the universal quantification over the set of computation runs as it occurs in the definition of $\models^+$ for the non-core formulae is never trivialised to that over an empty set. The sixth item states that for terminating events, the set of finite computation runs is contained in the set of finite computation sequences. In the case of non-termination, the set of finite computation runs is equated with the singleton set $\{\mathtt{fail}\}$. This latter equation provides for an easy way to ensure that agents have neither the opportunity nor the ability to successfully perform actions for which execution would take infinite time. The seventh item states that for an action α constituting a terminating event, the set of successor states reached by execution of the finite computation runs of α equals that reached by execution of the finite computation sequences of α. The eighth item states an analogous result with respect to abilities: the set of finite computation runs of which the agent is capable of performing equals that of the finite computation sequences for which the agent has the ability. The combination of the seventh and the eighth item indicates that we (at least partially) succeeded in singling out the set of *relevant* computation sequences in our definition of finite computation runs.

When interpreting the nondeterministic action constructor $+$ as formalised in Definition 4.22 and explained above, it indeed behaves as desired.

4.24. Proposition. *For all* $i \in \mathrm{A}$, $\alpha_1, \alpha_2 \in \mathrm{Ac}^+$ *and* $\varphi \in \mathrm{L}^+$ *we have:*

- $\models^+ \langle \mathrm{do}_i(\alpha_1 + \alpha_2)\rangle\varphi \leftrightarrow (\langle \mathrm{do}_i(\alpha_1)\rangle\varphi \wedge \langle \mathrm{do}_i(\alpha_2)\rangle\varphi)$
- $\models^+ \mathbf{A}_i(\alpha_1 + \alpha_2) \leftrightarrow (\mathbf{A}_i\alpha_1 \wedge \mathbf{A}_i\alpha_2)$

The validities that were found to characterise the ability, opportunity and result for confirmations and conditionally composed actions both in Chapter 3 and in Section 4.2, do also hold for $\models^+$. In addition, with regard to opportunity and result the sequential and repetitive composition behave as usual.

4.25. Proposition. *For all* $i \in \mathrm{A}$, $\alpha, \alpha_1, \alpha_2 \in \mathrm{Ac}^+$ *and* $\varphi, \psi \in \mathrm{L}^+$ *we have:*

1. $\models^+ \langle \mathrm{do}_i(\mathtt{confirm}\,\varphi)\rangle\psi \leftrightarrow (\varphi \wedge \psi)$
2. $\models^+ \langle \mathrm{do}_i(\alpha_1; \alpha_2)\rangle\psi \leftrightarrow \langle \mathrm{do}_i(\alpha_1)\rangle\langle \mathrm{do}_i(\alpha_2)\rangle\psi$
3. $\models^+ \langle \mathrm{do}_i(\mathtt{if}\,\varphi\,\mathtt{then}\,\alpha_1\,\mathtt{else}\,\alpha_2\,\mathtt{fi})\rangle\psi \leftrightarrow ((\varphi \wedge \langle \mathrm{do}_i(\alpha_1)\rangle\psi) \vee (\neg\varphi \wedge \langle \mathrm{do}_i(\alpha_2)\rangle\psi))$

4. $\models^+ \langle \mathrm{do}_i(\texttt{while}\,\varphi\,\texttt{do}\,\alpha\,\texttt{od})\rangle\psi \leftrightarrow ((\neg\varphi \wedge \psi) \vee (\varphi \wedge \langle \mathrm{do}_i(\alpha)\rangle\langle \mathrm{do}_i(\texttt{while}\,\varphi\,\texttt{do}\,\alpha\,\texttt{od})\rangle\psi))$
5. $\models^+ \mathbf{A}_i\texttt{confirm}\,\varphi \leftrightarrow \varphi$
6. $\models^+ \mathbf{A}_i\texttt{if}\,\varphi\,\texttt{then}\,\alpha_1\,\texttt{else}\,\alpha_2\,\texttt{fi} \leftrightarrow ((\varphi \wedge \mathbf{A}_i\alpha_1) \vee (\neg\varphi \wedge \mathbf{A}_i\alpha_2))$

The $[\mathrm{do}_i(\alpha)]$-operator is not normal when α is allowed to contain external nondeterministic actions. In particular, $[\mathrm{do}_i(\alpha)]$, while still preserving truth of the N-rule, no longer validates the K-axiom. The non-normality of $[\mathrm{do}_i(\alpha)]$ is due to the fact that this modality is no longer interpreted as a genuine necessity operator. In the interpretation of a formula $[\mathrm{do}_i(\alpha)]\varphi$ this is visible in the presence of an *existential*, rather than a *universal*, quantification over the finite computation runs of α.

4.26. Proposition. *For $i \in \mathrm{A}$ we have:*

- $[\mathrm{do}_i(\alpha)](\varphi \to \psi) \to ([\mathrm{do}_i(\alpha)]\varphi \to [\mathrm{do}_i(\alpha)]\psi)$ *is not for all* $\alpha \in \mathrm{Ac}^+$, $\varphi, \psi \in \mathrm{L}^+$, $\models^+$*-valid*
- $\models^+ \psi \Rightarrow \models^+ [\mathrm{do}_i(\alpha)]\psi$ *holds for all* $\alpha \in \mathrm{Ac}^+$, $\psi \in \mathrm{L}^+$

As expected and desired, in particular given the observations of Example 4.18, the validities found to characterise the agents' abilities for sequential and repetitive compositions for $\models^+$ differ from those found to characterise these abilities for $\models$.

4.27. Proposition. *For $i \in \mathrm{A}$, $\alpha, \alpha_1, \alpha_2, \alpha_3 \in \mathrm{Ac}^+$ and $\varphi \in \mathrm{L}^+$ we have:*

1. $\models^+ \mathbf{A}_i(\alpha_1;\alpha_2) \to \mathbf{A}_i\alpha_1 \wedge [\mathrm{do}_i(\alpha_1)]\mathbf{A}_i\alpha_2$
2. $\mathbf{A}_i\alpha_1 \wedge [\mathrm{do}_i(\alpha_1)]\mathbf{A}_i\alpha_2 \to \mathbf{A}_i(\alpha_1;\alpha_2)$ *is not for all* $\alpha_1, \alpha_2 \in \mathrm{Ac}^+$, $\models^+$*-valid*
3. $\models^+ \mathbf{A}_i((\alpha_1 + \alpha_2);\alpha_3) \leftrightarrow (\mathbf{A}_i(\alpha_1;\alpha_3) \wedge \mathbf{A}_i(\alpha_2;\alpha_3))$
4. $\models^+ \mathbf{A}_i(\alpha_1;(\alpha_2 + \alpha_3)) \leftrightarrow (\mathbf{A}_i(\alpha_1;\alpha_2) \wedge \mathbf{A}_i(\alpha_1;\alpha_3))$
5. $\models^+ \mathbf{A}_i\texttt{while}\,\varphi\,\texttt{do}\,\alpha\,\texttt{od} \leftrightarrow \neg\varphi \vee (\varphi \wedge \mathbf{A}_i\alpha;\texttt{while}\,\varphi\,\texttt{do}\,\alpha\,\texttt{od})$
6. $\models^+ \mathbf{A}_i\texttt{while}\,\varphi\,\texttt{do}\,\alpha\,\texttt{od} \to \neg\varphi \vee (\varphi \wedge \mathbf{A}_i\alpha \wedge [\mathrm{do}_i(\alpha)]\mathbf{A}_i\texttt{while}\,\varphi\,\texttt{do}\,\alpha\,\texttt{od})$
7. $(\varphi \wedge \mathbf{A}_i\alpha \wedge [\mathrm{do}_i(\alpha)]\mathbf{A}_i\texttt{while}\,\varphi\,\texttt{do}\,\alpha\,\texttt{od}) \to \mathbf{A}_i\texttt{while}\,\varphi\,\texttt{do}\,\alpha\,\texttt{od}$ *is not for all* $\varphi \in \mathrm{L}^+, \alpha \in \mathrm{Ac}^+$, $\models^+$*-valid*

For most of the items of Proposition 4.27 an analogous explanation can be given as the one provided for the corresponding items of Proposition 4.11. Items 1 through 4 formalise the (non-)validities that were explained to be desirable in Example 4.18. Item 5 concerns the universal characterisation of ability for repetitive composition: the equivalence given in this item is also valid for both $\models$ and $\models^{\oplus}$. Items 6 and 7 are again related to items 1 and 2.

4.3.2 External nondeterminism and pessimistic agents

At the beginning of Section 4.3 it was explained that for agents with a pessimistic view on their abilities for sequentially composed actions, the introduction of external nondeterministic actions would not necessarily have to lead to a complete change of the semantics for the non-core formulae. In particular it is no longer necessary to unravel actions in order to come up with an adequate semantics. The purpose of this section is to define a semantics for the language L^+ that is essentially a more or less straightforward extension of the semantics given for L. To this end we use elements of the semantics for Concurrent Propositional Dynamic Logic (CPDL) as it was given by Peleg [101].

In CPDL the class of actions of regular PDL is extended with an action constructor $\cap$, which models concurrency, i.e. the action $\alpha_1 \cap \alpha_2$ is intuitively interpreted as 'α_1 and α_2 in parallel'. The semantics that Peleg proposes to interpret the extended class of actions is such that $\langle \mathrm{do}_i(\alpha_1 \cap \alpha_2)\rangle\varphi \leftrightarrow (\langle \mathrm{do}_i(\alpha_1)\rangle\varphi \wedge \langle \mathrm{do}_i(\alpha_2)\rangle\varphi)$ comes out valid. Replacing $\cap$ by $+$ this is exactly the equivalence $(\Diamond)$ which was assumed to characterise external nondeterminism as far as results and opportunities are concerned[3]. Our idea therefore is on the one hand to adapt the semantics given by Peleg to make it suitable for our class of actions (which differs from the class of actions considered in regular PDL), and on the other hand to extend this semantics in order to interpret abilities in such a way that equivalence $(\blacklozenge)$ is validated. From a technical point of view there is no reason to propose an alternative to the semantics presented for external nondeterminism and optimistic agents. That is, a slight modification of the function c^+ would suffice to formalise external nondeterminism and pessimistic agents. The main reason for proposing another semantics here is that this alternative semantics is an elegant one, which is furthermore based on an approved one, viz. the semantics for CPDL as proposed by Peleg. The fundamental idea underlying Peleg's semantics is that actions are no longer seen as state-transitions but as transitions between states and sets of states. An event $\mathrm{do}_i(\alpha)$ corresponds to a set of pairs (s, U) with the intuitive interpretation that α can be executed by i in s, in parallel, to reach all states from U. In our interpretation of actions, the resulting set U of states formalises all states of affairs that an agent has to consider as possibly resulting from it executing α. The formula $\langle \mathrm{do}_i(\alpha)\rangle\varphi$ is now true in a state s of a model M if there is some set $U \subseteq \mathrm{S}$ such that execution of α in s by i leads to the set U, which is such that φ holds in all the states that are in U[4]. For technical reasons

[3]The resemblance between the concurrency operator $\cap$ and our external nondeterministic operator $+$, as far as opportunities and results are concerned, is not very surprising if one considers that Peleg explicitly states that in his approach concurrency is viewed as *the dual notion of nondeterminism*, where Peleg's notion of nondeterminism is identical to internal nondeterminism in our approach.

[4]Conceptually this interpretation of $\langle \mathrm{do}_i(\alpha)\rangle\varphi$ is highly similar to that of the belief formula $\mathbf{B}_i\varphi$ in the local reasoning approach of Fagin & Halpern [31]. The basic idea underlying the local reasoning approach is that of an agent of which the beliefs are spread out over various frames of reference. Formally

this set U is not defined to be the result of application of some function applied to i, α and s, but as the second element of a pair (s, U) for which a certain relation holds, i.e. we switch from a functional definition to a relational one. Abilities are still interpreted by a function applied to an action, an agent and a state and yielding a truth value.

4.28. DEFINITION. The binary relation $\models^{+,\mathbf{0}}$ between a formula from L^+ and a pair M, s consisting of a model M for L^+ and a state s in M is for the non-core formulae defined as follows:

$$\begin{array}{ll} \mathrm{M}, s \models^{+,\mathbf{0}} \langle \mathrm{do}_i(\alpha) \rangle \varphi & \Leftrightarrow \exists U((s,U) \in \mathtt{r}^{+,\mathbf{0}}(i,\alpha) \,\&\, \mathrm{M}, s' \models^{+,\mathbf{0}} \varphi \text{ for all } s' \in U) \\ \mathrm{M}, s \models^{+,\mathbf{0}} \mathbf{A}_i \alpha & \Leftrightarrow \mathtt{c}^{+,\mathbf{0}}(i,\alpha)(s) = \mathbf{1} \end{array}$$

where $\mathtt{r}^{+,\mathbf{0}}$ and $\mathtt{c}^{+,\mathbf{0}}$ are defined by:

$$\begin{array}{lcl} \mathtt{r}^{+,\mathbf{0}} & : & \mathrm{A} \times \mathrm{Ac}^+ \to \wp(\mathrm{S} \times \wp(\mathrm{S})) \\ \mathtt{r}^{+,\mathbf{0}}(i,a) & = & \{(s,\{s'\}) \mid \mathtt{r}_0(i,a)(s) = s'\} \\ \mathtt{r}^{+,\mathbf{0}}(i, \mathtt{confirm}\,\varphi) & = & \{(s,\{s\}) \mid \mathrm{M}, s \models^{+,\mathbf{0}} \varphi\} \\ \mathtt{r}^{+,\mathbf{0}}(i, \alpha_1; \alpha_2) & = & \mathtt{r}^{+,\mathbf{0}}(i,\alpha_1) \cdot \mathtt{r}^{+,\mathbf{0}}(i,\alpha_2) \\ \mathtt{r}^{+,\mathbf{0}}(i, \mathtt{if}\,\varphi\,\mathtt{then}\,\alpha_1\,\mathtt{else}\,\alpha_2\,\mathtt{fi}) & = & \mathtt{r}^{+,\mathbf{0}}(i, \mathtt{confirm}\,\varphi; \alpha_1) \cup \mathtt{r}^{+,\mathbf{0}}(i, \mathtt{confirm}\,\neg\varphi; \alpha_2) \\ \mathtt{r}^{+,\mathbf{0}}(i, \mathtt{while}\,\varphi\,\mathtt{do}\,\alpha\,\mathtt{od}) & = & \mathrm{LFP}(\mathrm{F}_{(i,\mathtt{confirm}\,\varphi;\alpha)}) \\ \mathtt{r}^{+,\mathbf{0}}(i, \alpha_1 + \alpha_2) & = & \{(s,U) \mid \exists U_1 \exists U_2((s,U_1) \in \mathtt{r}^{+,\mathbf{0}}(i,\alpha_1) \,\& \\ & & \qquad (s,U_2) \in \mathtt{r}^{+,\mathbf{0}}(i,\alpha_2) \,\&\, U = U_1 \cup U_2)\} \end{array}$$

$$\begin{array}{lcl} \mathtt{c}^{+,\mathbf{0}} & : & \mathrm{A} \times \mathrm{Ac}^+ \to \mathrm{S} \to \mathsf{bool} \\ \mathtt{c}^{+,\mathbf{0}}(i,a)(s) & = & \mathtt{c}_0(i,a)(s) \\ \mathtt{c}^{+,\mathbf{0}}(i, \mathtt{confirm}\,\varphi)(s) & = & \mathbf{1} \text{ iff } \mathrm{M}, s \models^{+,\mathbf{0}} \varphi \\ \mathtt{c}^{+,\mathbf{0}}(i, \alpha_1; \alpha_2)(s) & = & \mathbf{1} \text{ iff } \mathtt{c}^{+,\mathbf{0}}(i,\alpha_1)(s) = \mathbf{1} \,\&\, \exists U((s,U) \in \mathtt{r}^{+,\mathbf{0}}(i,\alpha_1) \,\& \\ & & \qquad \mathtt{c}^{+,\mathbf{0}}(i,\alpha_2)(s') = \mathbf{1} \text{ for all } s' \in U) \\ \mathtt{c}^{+,\mathbf{0}}(i, \mathtt{if}\,\varphi\,\mathtt{then}\,\alpha_1\,\mathtt{else}\,\alpha_2\,\mathtt{fi})(s) & = & \mathbf{1} \text{ iff } \mathtt{c}^{+,\mathbf{0}}(i, \mathtt{confirm}\,\varphi; \alpha_1)(s) = \mathbf{1} \text{ or} \\ & & \qquad \mathtt{c}^{+,\mathbf{0}}(i, \mathtt{confirm}\,\neg\varphi; \alpha_2)(s) = \mathbf{1} \\ \mathtt{c}^{+,\mathbf{0}}(i, \mathtt{while}\,\varphi\,\mathtt{do}\,\alpha\,\mathtt{od})(s) & = & \mathbf{1} \text{ iff } \exists U((s,U) \in \mathtt{r}^{+,\mathbf{0}}(i, \mathtt{while}\,\varphi\,\mathtt{do}\,\alpha\,\mathtt{od}) \,\& \\ & & \qquad (s,s') \in \mathtt{r}^{+,\mathbf{0}}(i,\alpha') \Rightarrow \mathtt{c}^{+,\mathbf{0}}(i,\alpha')(s) = \mathbf{1} \\ & & \qquad \text{for all } s' \in U, \alpha' \in \mathrm{CS}^+(\mathtt{while}\,\varphi\,\mathtt{do}\,\alpha\,\mathtt{od})) \\ \mathtt{c}^{+,\mathbf{0}}(i, \alpha_1 + \alpha_2)(s) & = & \mathbf{1} \text{ iff } \mathtt{c}^{+,\mathbf{0}}(i,\alpha_1)(s) = \mathbf{1} \,\&\mathtt{c}^{+,\mathbf{0}}(i,\alpha_2)(s) = \mathbf{1} \end{array}$$

where for $\mathrm{T}, \mathrm{T}_1, \mathrm{T}_2 \in \wp(\mathrm{S} \times \wp(\mathrm{S}))$

$$\begin{array}{lcl} \mathrm{T}_1 \cdot \mathrm{T}_2 & = & \{(s,U) \mid \exists s_1, U_1 \exists s_2, U_2 \ldots((s,\{s_1,s_2,\ldots\}) \in \mathrm{T}_1 \,\& \\ & & \qquad \forall k((s_k,U_k) \in \mathrm{T}_2 \,\&\, U = \textstyle\bigcup_k U_k)\} \\ \mathrm{F}_{(i,\mathtt{confirm}\,\varphi;\alpha)}(\mathrm{T}) & = & \mathtt{r}^{+,\mathbf{0}}(i, \mathtt{confirm}\,\neg\varphi) \cup \mathtt{r}^{+,\mathbf{0}}(i, \mathtt{confirm}\,\varphi; \alpha) \cdot \mathrm{T} \end{array}$$

and LFP yields the least fixed point of a function

---

this amounts to the agent considering various sets U of states, corresponding to its frames of reference, doxastically possible, while believing a formula if it is true in all the states of at least one such U.

The definition of $\mathtt{r}^{+,\mathbf{0}}$ for atomic actions, confirmations and conditional compositions is as usual and does not need any explanation; the same holds for the definition of $\mathtt{c}^{+,\mathbf{0}}$ for atomic actions, confirmations and conditional compositions. The operator $\cdot$ ties pairs (s_k, U_k). For example, if $(s, \{s_1, s_2\}) \in \mathtt{r}^{+,\mathbf{0}}(i, \alpha_1)$ and $\{(s_1, \{s_{11}, s_{12}\}), (s_2, \{s_{21}, s_{22}\})\} \subseteq \mathtt{r}^{+,\mathbf{0}}(i, \alpha_2)$ then $(s, \{s_{11}, s_{12}, s_{21}, s_{22}\}) \in \mathtt{r}^{+,\mathbf{0}}(i, \alpha_1) \cdot \mathtt{r}^{+,\mathbf{0}}(i, \alpha_2) = \mathtt{r}^{+,\mathbf{0}}(i, \alpha_1; \alpha_2)$. The definition of $\mathtt{r}^{+,\mathbf{0}}$ for while-loops is a little more elaborate than the corresponding definition of $\mathtt{r}$ as given in Chapter 3. The reason for this is that an action $\mathtt{while}\,\varphi\,\mathtt{do}\,\alpha\,\mathtt{od}$ may no longer be equated with its set of computation sequences, since this would force all possible ways of executing some action $\mathtt{while}\,\varphi\,\mathtt{do}\,\alpha_1 + \alpha_2\,\mathtt{od}$ to be of the same length, which leads to undesirable consequences.

4.29. Example. Consider a model M with a state s such that

- $\{(s, \{s_1\}), (s_2, \{s_{21}\})\} = \mathtt{r}^{+,\mathbf{0}}(i, a_1)$
- $\{(s, \{s_2\}), (s_2, \{s_{22}\})\} = \mathtt{r}^{+,\mathbf{0}}(i, a_2)$

and furthermore

- $\mathrm{M}, t \models^{+,\mathbf{0}} p$ for $t = s$ or $t = s_2$
- $\mathrm{M}, s_k \models^{+,\mathbf{0}} \neg p$ for $k = 1, 21, 22$

Intuitively one would expect $(s, \{s_1, s_{21}, s_{22}\}) \in \mathtt{r}^{+,\mathbf{0}}(i, \mathtt{while}\,p\,\mathtt{do}\,a_1 + a_2\,\mathtt{od})$. However, when defining $\mathtt{r}^{+,\mathbf{0}}(i, \mathtt{while}\,\varphi\,\mathtt{do}\,\alpha\,\mathtt{od}) = \bigcup_k \mathtt{r}^{+,\mathbf{0}}(i, (\mathtt{confirm}\,\varphi; \alpha)^k; \mathtt{confirm}\,\neg\varphi)$, we do not find $(s, \{s_1, s_{21}, s_{22}\}) \in \mathtt{r}^{+,\mathbf{0}}(i, \mathtt{while}\,p\,\mathtt{do}\,a_1 + a_2\,\mathtt{od})$. Due to the fact that the possible ways of executing $\mathtt{while}\,p\,\mathtt{do}\,a_1 + a_2\,\mathtt{od}$ are of different length, viz. a_1, $a_2; a_1$ and $a_2; a_2$, there is no k such that $(s, \{s_1, s_{21}, s_{22}\}) \in \mathtt{r}^{+,\mathbf{0}}(i, (\mathtt{confirm}\,p; (a_1 + a_2))^k; \mathtt{confirm}\,\neg p)$.

The counterintuitive situation encountered in Example 4.29 does not occur when employing the definition using fixed points as presented in 4.28. Since the operator $\cdot$ is monotone, i.e. whenever $\mathrm{T}_1 \subseteq \mathrm{T}_2$ also $\mathrm{T}_0 \cdot \mathrm{T}_1 \subseteq \mathrm{T}_0 \cdot \mathrm{T}_2$ for all T_0, $\mathrm{F}_{(i, \mathtt{confirm}\,\varphi; \alpha)}$ is also monotone, for every φ, α and i. By the Knaster-Tarski theorem [23, 124] this ensures that the least fixed point of $\mathrm{F}_{(i, \mathtt{confirm}\,\varphi; \alpha)}$ exists. Without further proof we state that this least fixed point can be computed as the limit of the following sequence of partial solutions:

- $\mathrm{F}_0 = \mathtt{r}^{+,\mathbf{0}}(i, \mathtt{confirm}\,\neg\varphi)$
- $\mathrm{F}_{k+1} = \mathrm{F}_0 \cup \mathtt{r}^{+,\mathbf{0}}(i, \mathtt{confirm}\,\varphi; \alpha) \cdot \mathrm{F}_k$
- $\mathrm{F}_\lambda = \bigcup_{\gamma < \lambda} \mathrm{F}_\gamma$ for a limit ordinal λ

For the model of Example 4.29 this translates to

- $\mathrm{F}_0 = \{(s_1, \{s_1\}), (s_{21}, \{s_{21}\}), (s_{22}, \{s_{22}\})\}$
- $\mathrm{F}_1 = \{(s_2, \{s_{21}, s_{22}\})\} \cup \mathrm{F}_0$
- $\mathrm{F}_2 = \{(s, \{s_1, s_{21}, s_{22}\})\} \cup \mathrm{F}_1$
- $\mathrm{F}_k = \mathrm{F}_2$ for all $k > 2$

Thus $\mathrm{r}^{+,\mathbf{0}}(i, \mathtt{while}\, p\, \mathtt{do}\, a_1 + a_2\, \mathtt{od}) = \mathrm{F}_2$, and hence in particular, $(s, \{s_1, s_{21}, s_{22}\}) \in \mathrm{r}^{+,\mathbf{0}}(i, \mathtt{while}\, p\, \mathtt{do}\, a_1 + a_2\, \mathtt{od})$, which is as intuitively desired.

The definition of $\mathrm{r}^{+,\mathbf{0}}$ for $\alpha_1 + \alpha_2$ formalises the idea that the action $\alpha_1 + \alpha_2$ corresponds to the 'sum' of α_1 and α_2, i.e. an agent has the opportunity to perform $\alpha_1 + \alpha_2$ iff it has both the opportunity to perform α_1 and α_2 while the result of $\alpha_1 + \alpha_2$ is determined by all formulae that are true both after executing α_1 and after α_2.

The definition of $\mathrm{c}^{+,\mathbf{0}}$ for both the sequential composition and the repetitive composition makes an essential use of the pessimistic view of agents on their abilities. In the definition of $\mathrm{c}^{+,\mathbf{0}}$ for $\alpha_1; \alpha_2$ it is demanded that execution of α_1 by agent i terminates in order to conclude that i has the ability to perform $\alpha_1; \alpha_2$. Furthermore, in all the states that together constitute the set of result states of execution of α_1 by i it has to be the case that i has the ability to perform α_2. As such, the definition of $\mathrm{c}^{+,\mathbf{0}}(i, \alpha_1; \alpha_2)$ can be seen as a generalisation of $\mathrm{c}(i, \alpha_1; \alpha_2)$, in which it is taken into account that execution of α_1 by i may lead to a set of result states rather than a single state. The definition of $\mathrm{c}^{+,\mathbf{0}}(i, \mathtt{while}\, \varphi\, \mathtt{do}\, \alpha\, \mathtt{od})$ is based on the idea that in the pessimistic approach while-loops are A-realisable. Hence for an agent to be able to perform the action $\mathtt{while}\, \varphi\, \mathtt{do}\, \alpha\, \mathtt{od}$ in a state s it has to be the case that some set U exists with $(s, U) \in \mathrm{r}^{+,\mathbf{0}}(i, \mathtt{while}\, \varphi\, \mathtt{do}\, \alpha\, \mathtt{od})$. Now the agent has to be able to execute all of the finite computation sequences of $\mathtt{while}\, \varphi\, \mathtt{do}\, \alpha\, \mathtt{od}$ of which execution leads to a state that is in U. For instance, in Example 4.29 it is indeed the case that execution of $\mathtt{while}\, p\, \mathtt{do}\, a_1 + a_2\, \mathtt{od}$ by i in s terminates. The set of finite computation sequences α' of $\mathtt{while}\, p\, \mathtt{do}\, a_1 + a_2\, \mathtt{od}$ that are such that $(s, \{s'\}) \in \mathrm{r}^{+,\mathbf{0}}(i, \alpha')$ for some $s' \in U$ are $(\mathtt{confirm}\, p; a_1); \mathtt{confirm}\, \neg p$, $(\mathtt{confirm}\, p; a_2); (\mathtt{confirm}\, p; a_1); \mathtt{confirm}\, \neg p$ and $(\mathtt{confirm}\, p; a_2); (\mathtt{confirm}\, p; a_2); \mathtt{confirm}\, \neg p$. Hence agent i is capable of performing $\mathtt{while}\, p\, \mathtt{do}\, a_1 + a_2\, \mathtt{od}$ iff it is able to perform all these three actions, which is indeed as intuitively desired.

As shown in Definition 4.28, it is perfectly possible to incorporate a formalisation of ability when using Peleg's approach as long as pessimistic agents are involved. However, it is hard to see how the same could be done for optimistic agents. As mentioned above, both in the definition of $\mathrm{c}^{+,\mathbf{0}}$ for sequential compositions and for repetitive compositions we essentially use some properties that are inherently due to the pessimistic view on ability, and these definitions are not easily adapted to an optimistic view on ability.

In the following proposition the semantics given in Definition 4.28 above is related to the ones based on the unravelling of actions. Furthermore, some desirable properties of the $\mathrm{r}^{+,\mathbf{0}}$-function are considered.

4.30. Proposition. *Let* M *be a model with state* s *and* U, U_1, U_2 *sets of states in* M. *Then for all* $i \in \mathrm{A}$, $\alpha \in \mathrm{Ac}^+$ *and* s' *in* M *we have:*

1. $(s, U) \in \mathrm{r}^{+,\mathbf{0}}(i, \alpha) \Rightarrow (s' \in U \Leftrightarrow \exists \alpha' \in \mathrm{CS}^+(\alpha)((s, \{s'\}) \in \mathrm{r}^{+,\mathbf{0}}(i, \alpha')))$

2. $(s,U) \in \mathtt{r}^{+,\mathbf{0}}(i,\alpha) \Rightarrow (\mathtt{c}^{+,\mathbf{0}}(i,\alpha)(s) = \mathbf{1} \Leftrightarrow (\mathtt{c}^{+,\mathbf{0}}(i,\alpha')(s) = \mathbf{1}$ *for all* $\alpha' \in \mathrm{CS}^+(\alpha)$ *with* $(s,\{s'\}) \in \mathtt{r}^{+,\mathbf{0}}(i,\alpha')$ *for some* $s' \in U))$
3. $(s,U) \in \mathtt{r}^{+,\mathbf{0}}(i,\alpha) \Rightarrow U \neq \emptyset \,\&\, \exists k \in \mathbb{N}(|U| \leq k)$
4. $(s,U_1) \in \mathtt{r}^{+,\mathbf{0}}(i,\alpha) \,\&\, (s,U_2) \in \mathtt{r}^{+,\mathbf{0}}(i,\alpha) \Rightarrow U_1 = U_2$

where $|U|$ *denotes the cardinality of* U.

The first item of Proposition 4.30 formalises the idea that every state s' in the set of states resulting from i executing an action α in s is reachable, and has in fact been reached, by i executing a particular finite computation sequence of α. The second item states that the notion of ability as formalised through the $\mathtt{c}^{+,\mathbf{0}}$ function may, under certain conditions, be seen as defined in terms of ability over finite computation sequences. The third item states two properties of the function $\mathtt{r}^{+,\mathbf{0}}$, the first of these being the fact that result sets of states are not empty. This property is essential for the definition of $\models^{+,\mathbf{0}}$ for dynamic formulae to be correct. For if a result set is allowed to be empty, this could lead to the formula $\langle \mathrm{do}_i(\mathtt{fail})\rangle\top$ being satisfiable. The second property formalised in the third item states that all actions α are, as Peleg calls it, *finitely branched*. This finiteness of branching is a consequence of the boundedness of the nondeterminism that we consider.

The semantics defined through the $\models^{+,\mathbf{0}}$ relation indeed combines the validities derived for the pessimistic approach in Chapter 3 with the equivalences $(\Diamond)$ and $(\blacklozenge)$ that characterise external nondeterminism. The following three propositions, the explanation of which has been given previously, summarise the various (in)validities that characterise results, opportunities and abilities for composite actions in the pessimistic approach.

4.31. Proposition. *For all* $i \in \mathrm{A}$, $\alpha, \alpha_1, \alpha_2 \in \mathrm{Ac}^+$ *and* $\varphi, \psi \in \mathrm{L}^+$ *we have:*

1. $\models^{+,\mathbf{0}} \langle \mathrm{do}_i(\mathtt{confirm}\,\varphi)\rangle\psi \leftrightarrow (\varphi \wedge \psi)$
2. $\models^{+,\mathbf{0}} \langle \mathrm{do}_i(\alpha_1;\alpha_2)\rangle\psi \leftrightarrow \langle \mathrm{do}_i(\alpha_1)\rangle\langle \mathrm{do}_i(\alpha_2)\rangle\psi$
3. $\models^{+,\mathbf{0}} \langle \mathrm{do}_i(\mathtt{if}\,\varphi\,\mathtt{then}\,\alpha_1\,\mathtt{else}\,\alpha_2\,\mathtt{fi})\rangle\psi \leftrightarrow ((\varphi \wedge \langle \mathrm{do}_i(\alpha_1)\rangle\psi) \vee (\neg\varphi \wedge \langle \mathrm{do}_i(\alpha_2)\rangle\psi))$
4. $\models^{+,\mathbf{0}} \langle \mathrm{do}_i(\mathtt{while}\,\varphi\,\mathtt{do}\,\alpha\,\mathtt{od})\rangle\psi \leftrightarrow ((\neg\varphi \wedge \psi) \vee (\varphi \wedge \langle \mathrm{do}_i(\alpha)\rangle\langle \mathrm{do}_i(\mathtt{while}\,\varphi\,\mathtt{do}\,\alpha\,\mathtt{od})\rangle\psi))$
5. $\models^{+,\mathbf{0}} \langle \mathrm{do}_i(\alpha_1 + \alpha_2)\rangle\varphi \leftrightarrow (\langle \mathrm{do}_i(\alpha_1)\rangle\varphi \wedge \langle \mathrm{do}_i(\alpha_2)\rangle\varphi)$

4.32. Proposition. *For* $i \in \mathrm{A}$ *we have:*

- $[\mathrm{do}_i(\alpha)](\varphi \to \psi) \to ([\mathrm{do}_i(\alpha)]\varphi \to [\mathrm{do}_i(\alpha)]\psi)$ *is not for all* $\alpha \in \mathrm{Ac}^+$, $\varphi, \psi \in \mathrm{L}^+$, $\models^{+,\mathbf{0}}$*-valid*
- $\models^{+,\mathbf{0}} \psi \Rightarrow \models^{+,\mathbf{0}} [\mathrm{do}_i(\alpha)]\psi$ *holds for all* $\alpha \in \mathrm{Ac}^+$, $\psi \in \mathrm{L}^+$

4.33. Proposition. *For all* $i \in \mathrm{A}$, $\alpha, \alpha_1, \alpha_2 \in \mathrm{Ac}^+$ *and* $\varphi \in \mathrm{L}^+$ *we have:*

1. $\models^{+,\mathbf{0}} \mathbf{A}_i\mathtt{confirm}\,\varphi \leftrightarrow \varphi$
2. $\models^{+,\mathbf{0}} \mathbf{A}_i\alpha_1;\alpha_2 \leftrightarrow \mathbf{A}_i\alpha_1 \wedge \langle \mathrm{do}_i(\alpha_1)\rangle\mathbf{A}_i\alpha_2$

3. $\models^{+,\mathbf{0}} \mathbf{A}_i \texttt{if}\ \varphi\ \texttt{then}\ \alpha_1\ \texttt{else}\ \alpha_2\ \texttt{fi} \leftrightarrow ((\varphi \wedge \mathbf{A}_i\alpha_1) \vee (\neg\varphi \wedge \mathbf{A}_i\alpha_2))$
4. $\models^{+,\mathbf{0}} \mathbf{A}_i \texttt{while}\ \varphi\ \texttt{do}\ \alpha\ \texttt{od} \leftrightarrow (\neg\varphi \vee (\varphi \wedge \mathbf{A}_i\alpha \wedge \langle \mathrm{do}_i(\alpha)\rangle \mathbf{A}_i \texttt{while}\ \varphi\ \texttt{do}\ \alpha\ \texttt{od}))$
5. $\models^{+,\mathbf{0}} \mathbf{A}_i(\alpha_1 + \alpha_2) \leftrightarrow (\mathbf{A}_i\alpha_1 \wedge \mathbf{A}_i\alpha_2)$

4.3.3 Practical possibility and forced choice

The problems that occurred with defining the Can-predicate and the Cannot-predicate in the presence of internal nondeterminism, do not arise in the presence of external nondeterminism. The semantics defined by $\models^{+}$ and $\models^{+,\mathbf{0}}$ is such that the aforementioned predicates display an intuitively acceptable behaviour when defined by the syntactical abbreviations proposed in Chapter 3.

4.34. Definition. For $\alpha \in \mathrm{Ac}^+$, $i \in \mathrm{A}$ and $\varphi \in \mathrm{L}^+$ we define:

- $\mathbf{PracPoss}_i(\alpha, \varphi) \triangleq \langle \mathrm{do}_i(\alpha)\rangle\varphi \wedge \mathbf{A}_i\alpha$
- $\mathbf{Can}_i(\alpha, \varphi) \triangleq \mathbf{K}_i\mathbf{PracPoss}_i(\alpha, \varphi)$
- $\mathbf{Cannot}_i(\alpha, \varphi) \triangleq \mathbf{K}_i\neg\mathbf{PracPoss}_i(\alpha, \varphi)$

That the behaviour of the Can-predicate and the Cannot-predicate is indeed intuitively acceptable when defined as above, is shown in the following proposition.

4.35. Proposition. *For $i \in \mathrm{A}$, $\alpha_1, \alpha_2 \in \mathrm{Ac}^+$ and $\varphi \in \mathrm{L}^+$ we have:*

1. $\models^x \mathbf{Can}_i(\alpha_1 + \alpha_2, \varphi) \leftrightarrow \mathbf{Can}_i(\alpha_1, \varphi) \wedge \mathbf{Can}_i(\alpha_2, \varphi)$
2. $\models^x \mathbf{Cannot}_i(\alpha_1, \varphi) \vee \mathbf{Cannot}_i(\alpha_2, \varphi) \rightarrow \mathbf{Cannot}_i(\alpha_1 + \alpha_2, \varphi)$
3. $\mathbf{Cannot}_i(\alpha_1 + \alpha_2, \varphi) \rightarrow \mathbf{Cannot}_i(\alpha_1, \varphi) \vee \mathbf{Cannot}_i(\alpha_2, \varphi)$ *is not for all* $\alpha_1, \alpha_2 \in \mathrm{Ac}^+, \varphi \in \mathrm{L}^+$, *$\models^x$-valid*

where $\models^x$ is either $\models^+$ or $\models^{+,\mathbf{0}}$

In Proposition 4.35 it is expressed in a nice way that the external choice is completely out of the control of the agent. In particular the fact that the implication in the last item of Proposition 4.35 is not valid seems reasonable: for should the agent conclude from the fact that it knows that $\alpha_1 + \alpha_2$ is either incorrect or infeasible to bring about φ, that it knows that either α_1 or α_2 is incorrect or infeasible to bring about φ, it seems to have some knowledge concerning the choice that the external environment makes or is going to make. But this would fiercely contradict our intuitive ideas on external nondeterminism as formulated in Section 4.1.

Using the Propositions 4.25 and 4.27, it is easily checked which of the validities given in Propositions 3.18 and 3.19 in Chapter 3 do also hold for $\models^{+}$. Since in Propositions 4.31 and 4.33 the same validities are found to hold for $\models^{+,\mathbf{0}}$ that were found to hold for $\models^{\mathbf{0}}$, the validities given in Section 3.3 to characterise the Can-predicate and the Cannot-predicate do also hold for $\models^{+,\mathbf{0}}$.

4.4 Summary and conclusions

In this chapter we dealt with both internal and external nondeterminism, by extending the class of action constructors with two new constructors, viz. $\_ \oplus \_$, representing the internal nondeterministic combination of two actions, and $\_ + \_$, denoting the external nondeterministic combination of two actions. When performing an internal nondeterministic action, the agent itself makes the (angelic) choice as to which action to perform, whereas when performing an external nondeterministic choice some unspecified external environment makes the (possibly demonic) choice. Due to the presence of abilities in our formal system the common approaches towards nondeterminism as they have been proposed in the literature, can not always and not straightforwardly be used. To deal with internal nondeterminism we proposed a semantics in which actions are unravelled into their elementary constituents. This semantics is applicable both for optimistic and pessimistic agents. The definitions of the Can-predicate and the Cannot-predicate as proposed in the previous chapter to model the practical possibilities of agents, have to be reconsidered in the presence of internal nondeterminism. When formalising external nondeterminism we distinguished between optimistic and pessimistic agents. For optimistic agents we proposed a semantics based on an elaborate and subtle unravelling of actions whereas for pessimistic agents we defined a semantics which extends the one proposed by Peleg to model concurrency in propositional dynamic logic. It turns out that the definition of known practical (im)possibility as proposed in the previous chapter is also applicable when external nondeterministic actions are considered.

4.4.1 Possible extensions

The most obvious and desirable extension of the formalisms presented in this chapter is a semantics in which internal nondeterminism and external nondeterminism are combined. Examples 4.2 and 4.18 indicate that the presence of abilities in our framework makes that the common approaches towards combinations of internal and external nondeterminism, some of which are mentioned below, cannot straightforwardly be applied. We feel that, although perhaps some clues may be found when combining the results of this chapter with some of the results found in the literature, a lot of research is necessary to come up with an adequate and intuitively acceptable account in which external and internal nondeterminism are combined in the presence of abilities, results and opportunities.

4.4.2 Bibliographical notes

The semantics based on unravelling given in this chapter was originally presented in [55]; the extension of Peleg's semantics used to deal with external nondeterminism and abilities was defined especially for this occasion.

Nondeterminism has been subject of intensive research in the theoretical computer science community (see for example [2, 11, 50]). Rather than providing a complete overview of the various formalisms proposed to model nondeterminism, we restrict ourselves to some approaches that, like our approach, are rooted in dynamic logic. It is important to remark that none of these approaches deals with ability: all are restricted to defining a semantics for dynamic formulae. The first of these approaches is dynamic logic itself: the action constructor $\_ + \_$, which, to make things easy, denotes the nondeterminism that we call internal, is an element of the set of standard action constructors for regular PDL [46]. The approach of Peleg [101], in which (internal) nondeterminism and concurrency are combined, and the variant proposed by Goldblatt [43], could be seen as formalisms in which internal and external nondeterminism are combined for results and opportunities. The semantics defined by Peleg, which is in a slightly modified form also used by Goldblatt, served as the foundation for our semantics of external nondeterminism for pessimistic agents. Another approach that we would like to mention is the one proposed by Meyer [93]. In the semantics of Meyer so-called '$\vee$-sets' and '$\wedge$-sets' are used. The $\wedge$-sets can be seen as corresponding to the union of the results sets as it is taken in the definition of $\mathtt{r}^{+,\mathbf{0}}(i, \alpha_1 + \alpha_2)$, the $\vee$-sets correspond to taking the union of $\mathtt{r}^{+,\mathbf{0}}(i, \alpha_1)$ and $\mathtt{r}^{+,\mathbf{0}}(i, \alpha_2)$. The approaches of Peleg, Goldblatt and Meyer combine internal and external nondeterminism for dynamic formulae, but are not as easily adapted to jointly deal with internal and external nondeterminism in the presence of abilities, since the counterintuitive situation formalised in Example 4.2 does occur in all three approaches.

4.5 Selected proofs

4.8. Proposition. *For all* $i \in \mathrm{A}$, $\alpha_1, \alpha_2 \in \mathrm{Ac}^{\oplus}$ *and* $\varphi \in \mathrm{L}^{\oplus}$ *we have:*

- $\models^{\oplus} \langle \mathrm{do}_i(\alpha_1 \oplus \alpha_2)\rangle\varphi \leftrightarrow (\langle \mathrm{do}_i(\alpha_1)\rangle\varphi \vee \langle \mathrm{do}_i(\alpha_2)\rangle\varphi)$
- $\models^{\oplus} \mathbf{A}_i(\alpha_1 \oplus \alpha_2) \leftrightarrow (\mathbf{A}_i\alpha_1 \vee \mathbf{A}_i\alpha_2)$

Proof: Let $\mathrm{M} \in \mathbf{M}^{\oplus}$ with state s, and $\alpha_1, \alpha_2 \in \mathrm{Ac}^{\oplus}$ and $\varphi \in \mathrm{L}^{\oplus}$ be arbitrary. We successively show both items.

$$
\begin{aligned}
&\mathrm{M}, s \models^{\oplus} \langle \mathrm{do}_i(\alpha_1 \oplus \alpha_2)\rangle\varphi \\
\Leftrightarrow\ &\exists \alpha' \in \mathrm{CS}^{\oplus}(\alpha_1 \oplus \alpha_2) \exists s' \in \mathrm{S}(\mathtt{r}^{\oplus}(i, \alpha')(s) = s' \,\&\, \mathrm{M}, s' \models^{\oplus} \varphi) \\
\Leftrightarrow\ &\exists \alpha' \in \mathrm{CS}^{\oplus}(\alpha_1) \cup \mathrm{CS}^{\oplus}(\alpha_2) \exists s' \in \mathrm{S}(\mathtt{r}^{\oplus}(i, \alpha')(s) = s' \,\&\, \mathrm{M}, s' \models^{\oplus} \varphi) \\
\Leftrightarrow\ &\exists \alpha' \in \mathrm{CS}^{\oplus}(\alpha_1) \exists s' \in \mathrm{S}(\mathtt{r}^{\oplus}(i, \alpha')(s) = s' \,\&\, \mathrm{M}, s' \models^{\oplus} \varphi) \text{ or} \\
&\exists \alpha' \in \mathrm{CS}^{\oplus}(\alpha_2) \exists s' \in \mathrm{S}(\mathtt{r}^{\oplus}(i, \alpha')(s) = s' \,\&\, \mathrm{M}, s' \models^{\oplus} \varphi) \\
\Leftrightarrow\ &\mathrm{M}, s \models^{\oplus} \langle \mathrm{do}_i(\alpha_1)\rangle\varphi \text{ or } \mathrm{M}, s \models^{\oplus} \langle \mathrm{do}_i(\alpha_2)\rangle\varphi \\
\Leftrightarrow\ &\mathrm{M}, s \models^{\oplus} \langle \mathrm{do}_i(\alpha_1)\rangle\varphi \vee \langle \mathrm{do}_i(\alpha_2)\rangle\varphi
\end{aligned}
$$

$\mathrm{M}, s \models^{\oplus} \mathbf{A}_i(\alpha_1 \oplus \alpha_2)$
$\Leftrightarrow \exists \alpha' \in \mathrm{CS}^{\oplus}(\alpha_1 \oplus \alpha_2)(\mathtt{c}^{\oplus}(i, \alpha')(s) = \mathbf{1})$
$\Leftrightarrow \exists \alpha' \in \mathrm{CS}^{\oplus}(\alpha_1) \cup \mathrm{CS}^{\oplus}(\alpha_2)(\mathtt{c}^{\oplus}(i, \alpha')(s) = \mathbf{1})$
$\Leftrightarrow \exists \alpha' \in \mathrm{CS}^{\oplus}(\alpha_1)(\mathtt{c}^{\oplus}(i, \alpha')(s) = \mathbf{1})$ or $\exists \alpha' \in \mathrm{CS}^{\oplus}(\alpha_2)(\mathtt{c}^{\oplus}(i, \alpha')(s) = \mathbf{1})$
$\Leftrightarrow \mathrm{M}, s \models^{\oplus} \mathbf{A}_i\alpha_1$ or $\mathrm{M}, s \models^{\oplus} \mathbf{A}_i\alpha_2$
$\Leftrightarrow \mathrm{M}, s \models^{\oplus} \mathbf{A}_i\alpha_1 \vee \mathbf{A}_i\alpha_2$
⊠

4.9. Proposition. *For all* $i \in \mathrm{A}$, $\alpha, \alpha_1, \alpha_2 \in \mathrm{Ac}^{\oplus}$ *and* $\varphi, \psi \in \mathrm{L}^{\oplus}$ *we have:*

1. $\models^{\oplus} \langle \mathrm{do}_i(\mathtt{confirm}\,\varphi)\rangle\psi \leftrightarrow (\varphi \wedge \psi)$
2. $\models^{\oplus} \langle \mathrm{do}_i(\alpha_1; \alpha_2)\rangle\psi \leftrightarrow \langle \mathrm{do}_i(\alpha_1)\rangle\langle \mathrm{do}_i(\alpha_2)\rangle\psi$
3. $\models^{\oplus} \langle \mathrm{do}_i(\mathtt{if}\,\varphi\,\mathtt{then}\,\alpha_1\,\mathtt{else}\,\alpha_2\,\mathtt{fi})\rangle\psi \leftrightarrow ((\varphi \wedge \langle \mathrm{do}_i(\alpha_1)\rangle\psi) \vee (\neg\varphi \wedge \langle \mathrm{do}_i(\alpha_2)\rangle\psi))$
4. $\models^{\oplus} \langle \mathrm{do}_i(\mathtt{while}\,\varphi\,\mathtt{do}\,\alpha\,\mathtt{od})\rangle\psi \leftrightarrow ((\neg\varphi \wedge \psi) \vee (\varphi \wedge \langle \mathrm{do}_i(\alpha)\rangle\langle \mathrm{do}_i(\mathtt{while}\,\varphi\,\mathtt{do}\,\alpha\,\mathtt{od})\rangle\psi))$
5. $\models^{\oplus} [\mathrm{do}_i(\alpha)](\varphi \to \psi) \to ([\mathrm{do}_i(\alpha)]\varphi \to [\mathrm{do}_i(\alpha)]\psi)$
6. $\models^{\oplus} \psi \Rightarrow \models^{\oplus} [\mathrm{do}_i(\alpha)]\psi$

Proof: The first item follows since $\mathrm{CS}^{\oplus}(\mathtt{confirm}\,\varphi) = \mathtt{confirm}\,\varphi$, and $\mathtt{r}^{\oplus}(i, \mathtt{confirm}\,\varphi)(s)$ is defined as is $\mathtt{r}(i, \mathtt{confirm}\,\varphi)(s)$. Items 3 and 6 are obvious, and item 4 is proved in a similar way as item 2. Here we show items 2 and 5. Let $\mathrm{M} \in \mathbf{M}^{\oplus}$ with state s, and $\alpha, \alpha_1, \alpha_2 \in \mathrm{Ac}^{\oplus}$ and $\psi \in \mathrm{L}^{\oplus}$ be arbitrary.

2. $\mathrm{M}, s \models^{\oplus} \langle \mathrm{do}_i(\alpha_1; \alpha_2)\rangle\psi$
$\Leftrightarrow \exists \alpha' \in \mathrm{CS}^{\oplus}(\alpha_1; \alpha_2) \exists s' \in \mathrm{S}(\mathtt{r}^{\oplus}(i, \alpha')(s) = s' \,\&\, \mathrm{M}, s' \models^{\oplus} \psi)$
$\Leftrightarrow \exists \alpha'_1 \in \mathrm{CS}^{\oplus}(\alpha_1) \exists \alpha'_2 \in \mathrm{CS}^{\oplus}(\alpha_2) \exists s' \in \mathrm{S}(\mathtt{r}^{\oplus}(i, \alpha'_1; \alpha'_2)(s) = s' \,\&\, \mathrm{M}, s' \models^{\oplus} \psi)$
$\Leftrightarrow \exists \alpha'_1 \in \mathrm{CS}^{\oplus}(\alpha_1) \exists \alpha'_2 \in \mathrm{CS}^{\oplus}(\alpha_2) \exists s' \in \mathrm{S} \exists s'' \in \mathrm{S}(\mathtt{r}^{\oplus}(i, \alpha'_1)(s) = s'' \,\&$
$\mathtt{r}^{\oplus}(i, \alpha'_2)(s'') = s' \,\&\, \mathrm{M}, s' \models^{\oplus} \psi)$
$\Leftrightarrow \exists \alpha'_1 \in \mathrm{CS}^{\oplus}(\alpha_1) \exists s'' \in \mathrm{S}(\mathtt{r}^{\oplus}(i, \alpha'_1)(s) = s'' \,\&$
$\exists \alpha'_2 \in \mathrm{CS}^{\oplus}(\alpha_2) \exists s' \in \mathrm{S}(\mathtt{r}^{\oplus}(i, \alpha'_2)(s'') = s' \,\&\, \mathrm{M}, s' \models^{\oplus} \psi))$
$\Leftrightarrow \exists \alpha'_1 \in \mathrm{CS}^{\oplus}(\alpha_1) \exists s'' \in \mathrm{S}(\mathtt{r}^{\oplus}(i, \alpha'_1)(s) = s'' \,\&\, \mathrm{M}, s'' \models^{\oplus} \langle \mathrm{do}_i(\alpha_2)\rangle\psi)$
$\Leftrightarrow \mathrm{M}, s \models^{\oplus} \langle \mathrm{do}_i(\alpha_1)\rangle\langle \mathrm{do}_i(\alpha_2)\rangle\psi$

5. $\mathrm{M}, s \models^{\oplus} [\mathrm{do}_i(\alpha)](\varphi \to \psi)$
$\Leftrightarrow \mathrm{M}, s \models^{\oplus} \neg\langle \mathrm{do}_i(\alpha)\rangle\neg(\varphi \to \psi)$
$\Leftrightarrow \mathsf{not}(\mathrm{M}, s \models^{\oplus} \langle \mathrm{do}_i(\alpha)\rangle\neg(\varphi \to \psi))$
$\Leftrightarrow \mathsf{not}(\exists \alpha' \in \mathrm{CS}^{\oplus}(\alpha) \exists s' \in \mathrm{S}(\mathtt{r}^{\oplus}(i, \alpha')(s) = s' \,\&\, \mathsf{not}(\mathrm{M}, s' \models^{\oplus} (\varphi \to \psi))))$
$\Leftrightarrow \forall \alpha' \in \mathrm{CS}^{\oplus}(\alpha) \forall s' \in \mathrm{S}(\mathtt{r}^{\oplus}(i, \alpha')(s) = s' \Rightarrow \mathrm{M}, s' \models^{\oplus} (\varphi \to \psi))$ $(*)$

In a similar way we can show that $\mathrm{M}, s \models^{\oplus} [\mathrm{do}_i(\alpha)]\varphi$ iff $\forall \alpha' \in \mathrm{CS}^{\oplus}(\alpha) \forall s' \in \mathrm{S}(\mathtt{r}^{\oplus}(i, \alpha')(s) = s' \Rightarrow \mathrm{M}, s' \models^{\oplus} \varphi)$ $(**)$. Combining $*$ and $**$ leads to $\forall \alpha' \in \mathrm{CS}^{\oplus}(\alpha) \forall s' \in \mathrm{S}(\mathtt{r}^{\oplus}(i, \alpha')(s) = s' \Rightarrow \mathrm{M}, s' \models^{\oplus} \psi)$, which is necessary and sufficient to conclude that $\mathrm{M}, s \models^{\oplus} [\mathrm{do}_i(\alpha)]\psi$. Hence, if $\mathrm{M}, s \models^{\oplus} [\mathrm{do}_i(\alpha)](\varphi \to \psi)$ then if $\mathrm{M}, s \models^{\oplus} [\mathrm{do}_i(\alpha)]\varphi$ also $\mathrm{M}, s \models^{\oplus} [\mathrm{do}_i(\alpha)]\psi$, which suffices to conclude item 5.

⊠

4.11. Proposition. *For* $i \in \mathrm{A}$, $\alpha, \alpha_1, \alpha_2, \alpha_3 \in \mathrm{Ac}^{\oplus}$ *and* $\varphi \in \mathrm{L}^{\oplus}$ *we have:*

1. $\mathbf{A}_i\alpha_1;\alpha_2 \to [\mathrm{do}_i(\alpha_1)]\mathbf{A}_i\alpha_2$ *is not for all* $\alpha_1, \alpha_2 \in \mathrm{Ac}^{\oplus}$, $\models^{\oplus}$*-valid*
2. $\models^{\oplus} \mathbf{A}_i\alpha_1 \wedge [\mathrm{do}_i(\alpha_1)]\mathbf{A}_i\alpha_2 \to \mathbf{A}_i\alpha_1;\alpha_2$
3. $\models^{\oplus} \mathbf{A}_i(\alpha_1 \oplus \alpha_2);\alpha_3 \leftrightarrow \mathbf{A}_i(\alpha_1;\alpha_3) \vee \mathbf{A}_i(\alpha_2;\alpha_3)$
4. $\models^{\oplus} \mathbf{A}_i\alpha_1;(\alpha_2 \oplus \alpha_3) \leftrightarrow \mathbf{A}_i(\alpha_1;\alpha_2) \vee \mathbf{A}_i(\alpha_1;\alpha_3)$
5. $\models^{\oplus} \mathbf{A}_i\mathtt{while}\,\varphi\,\mathtt{do}\,\alpha\,\mathtt{od} \leftrightarrow \neg\varphi \vee (\varphi \wedge \mathbf{A}_i\alpha;\mathtt{while}\,\varphi\,\mathtt{do}\,\alpha\,\mathtt{od})$
6. $\mathbf{A}_i\mathtt{while}\,\varphi\,\mathtt{do}\,\alpha\,\mathtt{od} \to (\neg\varphi \vee [\mathrm{do}_i(\alpha)]\mathbf{A}_i\mathtt{while}\,\varphi\,\mathtt{do}\,\alpha\,\mathtt{od})$ *is not for all* $\varphi \in \mathrm{L}^{\oplus}, \alpha \in \mathrm{Ac}^{\oplus}$, $\models^{\oplus}$*-valid*
7. $\models^{\oplus} (\neg\varphi \vee (\varphi \wedge \mathbf{A}_i\alpha \wedge [\mathrm{do}_i(\alpha)]\mathbf{A}_i\mathtt{while}\,\varphi\,\mathtt{do}\,\alpha\,\mathtt{od})) \to \mathbf{A}_i\mathtt{while}\,\varphi\,\mathtt{do}\,\alpha\,\mathtt{od}$

Proof: We show the first two items; items 3, 4 and 5 are obvious and 6 and 7 are shown in a similar way as 1 and 2.

1. Assume that $i \in \mathrm{A}$ and $a_k \in \mathrm{At}$ for $k = 1, 2, 3$. Let $\mathrm{M} = \langle \mathrm{S}, \pi, \mathrm{R}, \mathtt{r}_0, \mathtt{c}_0 \rangle$ be such that
   - $\mathrm{S} = \{s_0, s_1, s_2\}$
   - π and R are arbitrary
   - $\mathtt{r}_0(i, a_k)(s_0) = s_k$ for $k = 1, 2$
   - $\mathtt{c}_0(i, a_1)(s_0) = \mathbf{1} = \mathtt{c}_0(i, a_3)(s_1), \mathtt{c}_0(i, a_2)(s_0) = \mathbf{0} = \mathtt{c}_0(i, a_3)(s_2)$

   For this model it holds that
   - $\mathrm{M}, s_0 \models^{\oplus} \mathbf{A}_i(a_1 \oplus a_2);a_3$
   - $\mathrm{M}, s_0 \not\models^{\oplus} [\mathrm{do}_i(a_1 \oplus a_2)]\mathbf{A}_i a_3$

   which suffices to conclude that M is a counterexample to item 1.
2. Let $\mathrm{M} \in \mathbf{M}^{\oplus}$ with state s and $\alpha_1, \alpha_2 \in \mathrm{Ac}^{\oplus}$ be arbitrary.

   $\mathrm{M}, s \models^{\oplus} \mathbf{A}_i\alpha_1 \wedge [\mathrm{do}_i(\alpha_1)]\mathbf{A}_i\alpha_2$
   $\Leftrightarrow \exists \alpha_1' \in \mathrm{CS}^{\oplus}(\alpha_1)(\mathtt{c}^{\oplus}(i, \alpha_1')(s) = \mathbf{1})\,\&$
   $\forall \alpha_1' \in \mathrm{CS}^{\oplus}(\alpha_1)\forall s' \in \mathrm{S}(\mathtt{r}^{\oplus}(i, \alpha_1')(s) = s' \Rightarrow \mathrm{M}, s' \models^{\oplus} \mathbf{A}_i\alpha_2)$
   $\Leftrightarrow \exists \alpha_1' \in \mathrm{CS}^{\oplus}(\alpha_1)(\mathtt{c}^{\oplus}(i, \alpha_1')(s) = \mathbf{1})\,\&$
   $\forall \alpha_1' \in \mathrm{CS}^{\oplus}(\alpha_1)\forall s' \in \mathrm{S}(\mathtt{r}^{\oplus}(i, \alpha_1')(s) = s' \Rightarrow \exists \alpha_2' \in \mathrm{CS}^{\oplus}(\alpha_2)(\mathtt{c}^{\oplus}(i, \alpha_2')(s') = \mathbf{1}))$

   Now let $\alpha_1' \in \mathrm{CS}^{\oplus}(\alpha_1)$ be such that $\mathtt{c}^{\oplus}(i, \alpha_1')(s) = \mathbf{1}$. We distinguish two cases:
   - $\mathtt{r}^{\oplus}(i, \alpha_1')(s) = \emptyset$. Take some $\alpha_2' \in \mathrm{CS}^{\oplus}(\alpha_2)$; this is possible since $\mathrm{CS}^{\oplus}(\alpha) \neq \emptyset$ for all $\alpha \in \mathrm{Ac}^{\oplus}$. Then it holds that $\alpha_1';\alpha_2' \in \mathrm{CS}^{\oplus}(\alpha_1;\alpha_2)$ and $\mathtt{c}^{\oplus}(i, \alpha_1';\alpha_2')(s) = \mathbf{1}$, and hence $\mathrm{M}, s \models^{\oplus} \mathbf{A}_i\alpha_1;\alpha_2$.
   - $\mathtt{r}^{\oplus}(i, \alpha_1')(s) = s'$, for some $s' \in \mathrm{S}$. It then follows that $\mathtt{c}^{\oplus}(i, \alpha_2')(s') = \mathbf{1}$ for some $\alpha_2' \in \mathrm{CS}^{\oplus}(\alpha_2)$. Hence $\mathtt{c}^{\oplus}(i, \alpha_1';\alpha_2')(s) = \mathbf{1}$ and since $\alpha_1';\alpha_2' \in \mathrm{CS}^{\oplus}(\alpha_1;\alpha_2)$ we conclude that $\mathrm{M}, s \models^{\oplus} \mathbf{A}_i\alpha_1;\alpha_2$.

   Since in both cases $\mathrm{M}, s \models^{\oplus} \mathbf{A}_i\alpha_1;\alpha_2$ we conclude that item 2 holds.

⊠

4.16. Proposition. *For $i \in \mathrm{A}$ we have:*

- $\mathbf{Can}_i(\alpha, \varphi) \wedge \mathbf{Can}_i(\alpha, \neg\varphi)$ *is $\models^{\oplus}$-satisfiable for certain $\alpha \in \mathrm{Ac}^{\oplus}$, $\varphi \in \mathrm{L}^{\oplus}$*
- $\mathbf{Can}_i(\alpha, \varphi) \wedge \mathbf{Can}_i(\alpha, \psi) \rightarrow \mathbf{Can}_i(\alpha, \varphi \wedge \psi)$ *is not for all $\alpha \in \mathrm{Ac}^{\oplus}$, $\psi \in \mathrm{L}^{\oplus}$, $\models^{\oplus}$-valid*

Proof: Both items are shown by the model $\mathrm{M} = \langle \mathrm{S}, \pi, \mathrm{R}, \mathtt{c}_0, \mathtt{c}_0 \rangle$ which is for $p \in \Pi$, $i \in \mathrm{A}$ and $a_1, a_2 \in \mathrm{At}$ such that

- $\mathrm{S} = \{s, s_1, s_2\}$
- $\pi(p, s_1) = \mathbf{1}$, $\pi(p, s_2) = \mathbf{0}$
- $\mathrm{R}(i) = \{(s, s), (s_1, s_1), (s_2, s_2)\}$
- $\mathtt{r}_0(i, a_k)(s) = s_k$ for $k = 1, 2$
- $\mathtt{c}_0(i, a_1)(s) = \mathbf{1} = \mathtt{c}_0(i, a_2)(s)$

For this model it holds that

- $\mathrm{M}, s \models^{\oplus} \mathbf{Can}_i(a_1 \oplus a_2, p)$ since $\mathrm{M}, s \models^{\oplus} \langle \mathrm{do}_i(a_1) \rangle p \wedge \mathbf{A}_i a_1$
- $\mathrm{M}, s \models^{\oplus} \mathbf{Can}_i(a_1 \oplus a_2, \neg p)$ since $\mathrm{M}, s \models^{\oplus} \langle \mathrm{do}_i(a_2) \rangle \neg p \wedge \mathbf{A}_i a_2$
- $\mathrm{M}, s \not\models^{\oplus} \mathbf{Can}_i(a_1 \oplus a_2, p \wedge \neg p)$ since there is no $\alpha' \in \mathrm{CS}^{\oplus}(a_1 \oplus a_2)$ such that $\mathrm{M}, s \models^{\oplus} \langle \mathrm{do}_i(\alpha') \rangle (p \wedge \neg p) \wedge \mathbf{A}_i \alpha'$

These three properties of M suffice to conclude that it shows both items of Proposition 4.16.

⊠

4.23. Proposition. *For all $\mathrm{M} \in \mathbf{M}^+$, s, s' in M, $i \in \mathrm{A}$, $\alpha, \alpha_1, \alpha_2 \in \mathrm{Ac}^+$ and $\varphi \in \mathrm{L}^+$:*

1. $\mathrm{Ter}_{\mathrm{M}}(i, \alpha_1; \alpha_2, s) = \mathbf{1} \Rightarrow \mathrm{Ter}_{\mathrm{M}}(i, \alpha_1, s) = \mathbf{1}$ &
   $\forall s'(\exists \alpha_1' \in \mathrm{CS}^+(\alpha_1)(s' = \mathtt{r}^+(i, \alpha_1')(s)) \Rightarrow \mathrm{Ter}_{\mathrm{M}}(i, \alpha_2, s') = \mathbf{1})$
2. $\mathrm{Ter}_{\mathrm{M}}(i, \mathtt{if}\ \varphi\ \mathtt{then}\ \alpha_1\ \mathtt{else}\ \alpha_2\ \mathtt{fi}, s) = \mathbf{1} \Rightarrow$
   $\mathrm{Ter}_{\mathrm{M}}(i, \mathtt{confirm}\ \varphi; \alpha_1, s) = \mathbf{1}$ & $\mathrm{Ter}_{\mathrm{M}}(i, \mathtt{confirm}\ \neg\varphi; \alpha_2, s) = \mathbf{1}$
3. $\mathrm{Ter}_{\mathrm{M}}(i, \mathtt{while}\ \varphi\ \mathtt{do}\ \alpha\ \mathtt{od}, s) = \mathbf{1} \Rightarrow$
   $\forall \alpha'' \in \mathrm{Ac}_{\mathrm{b}}^+(\exists \alpha' \in \mathrm{CS}^+(\mathtt{while}\ \varphi\ \mathtt{do}\ \alpha\ \mathtt{od})(\mathrm{Prefix}^+(\alpha'', \alpha')) \Rightarrow$
   $(\mathtt{r}^+(i, \alpha'')(s) = s' \Rightarrow \mathrm{Ter}_{\mathrm{M}}(i, \mathtt{confirm}\ \varphi; \alpha, s') = \mathbf{1}))$
4. $\mathrm{Ter}_{\mathrm{M}}(i, \alpha_1 + \alpha_2, s) = \mathbf{1} \Rightarrow \mathrm{Ter}_{\mathrm{M}}(i, \alpha_1, s) = \mathbf{1}$ & $\mathrm{Ter}_{\mathrm{M}}(i, \alpha_2, s) = \mathbf{1}$
5. $\mathrm{CS}^+(\alpha) \neq \emptyset$ *and* $\mathrm{CR}_{\mathrm{M}}^+(i, \alpha, s) \neq \emptyset$
6. $\mathrm{Ter}_{\mathrm{M}}(i, \alpha, s) = \mathbf{1} \Rightarrow \mathrm{CR}_{\mathrm{M}}^+(i, \alpha, s) \subseteq \mathrm{CS}^+(\alpha)$
7. $\mathrm{Ter}_{\mathrm{M}}(i, \alpha, s) = \mathbf{1} \Rightarrow \forall \alpha' \in \mathrm{Ac}_{\mathrm{b}}^+ \forall s'(\alpha' \in \mathrm{CS}^+(\alpha)$ & $\mathtt{r}^+(i, \alpha')(s) = s' \Leftrightarrow$
   $\alpha' \in \mathrm{CR}_{\mathrm{M}}^+(i, \alpha, s)$ & $\mathtt{r}^+(i, \alpha')(s) = s')$
8. $\mathrm{Ter}_{\mathrm{M}}(i, \alpha, s) = \mathbf{1} \Rightarrow \forall \alpha' \in \mathrm{Ac}_{\mathrm{b}}^+(\alpha' \in \mathrm{CS}^+(\alpha)$ & $\mathtt{c}^+(i, \alpha')(s) = \mathbf{1} \Leftrightarrow$
   $\alpha' \in \mathrm{CR}_{\mathrm{M}}^+(i, \alpha, s)$ & $\mathtt{c}^+(i, \alpha')(s) = \mathbf{1})$

Proof: The second and fourth item are easily shown and left to the reader. The proofs of items 7 and 8 are rather elaborate and fairly complicated and not given here. The interested reader is kindly referred to [53] where these proofs are presented in all detail.

Items 5 and 6 are shown by induction on the structure of α, using items 1 through 4 of this proposition in the proof of item 6. Here we show the first and the third item.

1. Firstly we show that if $\mathrm{Ter}_{\mathrm{M}}(i, \alpha_1, s) = \mathbf{0}$ then $\mathrm{Ter}_{\mathrm{M}}(i, \alpha_1; \alpha_2, s) = \mathbf{0}$, and secondly that if for some $s' = \mathtt{r}^+(i, \alpha_1')(s)$ with $\alpha_1' \in \mathrm{CS}^+(\alpha_1)$, it holds that $\mathrm{Ter}_{\mathrm{M}}(i, \alpha_2, s') = \mathbf{0}$, then $\mathrm{Ter}_{\mathrm{M}}(i, \alpha_1; \alpha_2, s) = \mathbf{0}$. With this the contraposition of the implication given above — and thereby the implication itself — is proved.
   - Assume that $\mathrm{Ter}_{\mathrm{M}}(i, \alpha_1, s) = \mathbf{0}$. We show that $\mathrm{Ter}_{\mathrm{M}}(i, \alpha_1; \alpha_2, s) = \mathbf{0}$ by showing that for all $k \in \mathbb{N}$, an $l \geq k$ and actions $\alpha',\ \alpha''$ exist, such that $\alpha'' = |\alpha'|_l^+,\ \alpha' \in \mathrm{CS}^+(\alpha_1; \alpha_2)$ and $\mathtt{r}^+(i, \alpha'')(s) \neq \emptyset$. Let $k \in \mathbb{N}$. Since $\mathrm{Ter}_{\mathrm{M}}(i, \alpha_1, s) = \mathbf{0}$ some $l \geq k,\ \beta',\ \beta''$ exist such that $\beta'' = |\beta'|_l^+,\ \beta' \in \mathrm{CS}^+(\alpha_1)$ and $\mathtt{r}^+(i, \beta'')(s) \neq \emptyset$. Now take some $\alpha_2' \in \mathrm{CS}^+(\alpha_2)$; such an α_2' exists since $\mathrm{CS}^+(\alpha_2) \neq \emptyset$. We have that $\beta'' = |\beta'; \alpha_2'|_l^+,\ l \geq k,\ \beta'; \alpha_2' \in \mathrm{CS}^+(\alpha_1; \alpha_2)$ and $\mathtt{r}^+(i, \beta'')(s) \neq \emptyset$. Since k was chosen arbitrarily in $\mathbb{N}$ this suffices to conclude that $\mathrm{Ter}_{\mathrm{M}}(i, \alpha_1; \alpha_2, s) = \mathbf{0}$.
   - Assume that $s' = \mathtt{r}^+(i, \alpha_1')(s)$ with $\alpha_1' \in \mathrm{CS}^+(\alpha_1)$ is such that $\mathrm{Ter}_{\mathrm{M}}(i, \alpha_2, s') = \mathbf{0}$. As in the previous case we show that $\mathrm{Ter}_{\mathrm{M}}(i, \alpha_1; \alpha_2, s) = \mathbf{0}$ by showing that for all $k \in \mathbb{N}$, an $l \geq k$ and actions $\alpha',\ \alpha''$ exist, such that $\alpha'' = |\alpha'|_l^+,\ \alpha' \in \mathrm{CS}^+(\alpha_1; \alpha_2)$ and $\mathtt{r}^+(i, \alpha'')(s) \neq \emptyset$. Let $k \in \mathbb{N}$ be arbitrary. Since $\mathrm{Ter}_{\mathrm{M}}(i, \alpha_2, s') = \mathbf{0}$ some $l \geq k,\ \gamma',\ \gamma''$ exist such that $\gamma'' = |\gamma'|_l^+,\ \gamma' \in \mathrm{CS}^+(\alpha_2)$ and $\mathtt{r}^+(i, \gamma'')(s') \neq \emptyset$. It now holds that $\alpha_1'; \gamma' \in \mathrm{CS}^+(\alpha_1; \alpha_2),\ |\alpha_1'; \gamma''|^+ = l' \geq l \geq k$, and $\mathtt{r}^+(i, \alpha_1'; \gamma'')(s) \neq \emptyset$. Since k was chosen arbitrarily in $\mathbb{N}$ it follows that $\mathrm{Ter}_{\mathrm{M}}(i, \alpha_1; \alpha_2, s) = \mathbf{0}$.

   Since in both cases $\mathrm{Ter}_{\mathrm{M}}(i, \alpha_1; \alpha_2, s) = \mathbf{0}$ we conclude that (the contraposition of) item 1 holds.

3. Assume that $\mathrm{Ter}_{\mathrm{M}}(i, \mathtt{while}\,\varphi\,\mathtt{do}\,\alpha\,\mathtt{od}, s) = \mathbf{1}$. Let $k \in \mathbb{N}$ be such that $\forall l \geq k \forall \alpha' \forall \alpha''(\alpha'' = |\alpha'|_l^+ \,\&\, \alpha' \in \mathrm{CS}^+(\mathtt{while}\,\varphi\,\mathtt{do}\,\alpha\,\mathtt{od}) \Rightarrow \mathtt{r}^+(i, \alpha'')(s) = \emptyset)$. Assume that $s',\ \alpha''$ and α' are such that $\mathrm{Prefix}^+(\alpha'', \alpha'),\ \alpha' \in \mathrm{CS}^+(\mathtt{while}\,\varphi\,\mathtt{do}\,\alpha\,\mathtt{od})$, and $\mathtt{r}^+(i, \alpha'')(s) = s'$. Assume towards a contradiction that $\mathrm{Ter}_{\mathrm{M}}(i, \mathtt{confirm}\,\varphi; \alpha, s') = \mathbf{0}$. Let $l \geq k$ and $\beta',\ \beta''$ be such that $\beta'' = |\beta'|_l^+,\ \beta' \in \mathrm{CS}^+(\mathtt{confirm}\,\varphi; \alpha)$ and $\mathtt{r}^+(i, \beta'')(s') \neq \emptyset$. Note that, since $\mathrm{Ter}_{\mathrm{M}}(i, \mathtt{confirm}\,\varphi; \alpha, s') = \mathbf{0}$, it holds that $\mathrm{M}, s' \models^+ \varphi$. This implies that α'', which is such that $\mathtt{r}^+(i, \alpha'')(s) = s'$, may not end in a confirmation for $\neg\varphi$. Hence this α'' has any of the following three forms:
   - $\alpha'' = \mathtt{confirm}\,\varphi$
   - $\alpha'' = (\mathtt{confirm}\,\varphi; \alpha_1'); \ldots; (\mathtt{confirm}\,\varphi; \alpha_l')$, for $l \geq 1$
   - $\alpha'' = (\mathtt{confirm}\,\varphi; \alpha_1'); \ldots; (\mathtt{confirm}\,\varphi; \alpha_l'); \mathtt{confirm}\,\varphi$, for $l \geq 1$

   where $\alpha_1', \ldots, \alpha_l'$ are in $\mathrm{CS}^+(\alpha)$. Define the action γ in any of the three cases given above as follows:
   - $\gamma = \beta''$
   - $\gamma = (\mathtt{confirm}\,\varphi; \alpha_1'); \ldots; (\mathtt{confirm}\,\varphi; \alpha_l'); \beta''$
   - $\gamma = (\mathtt{confirm}\,\varphi; \alpha_1'); \ldots; (\mathtt{confirm}\,\varphi; \alpha_l'); \beta''$

Now $\gamma = |\gamma;\mathtt{confirm}\,\neg\varphi|^+_m$ with $m \geq l \geq k$, $(\gamma;\mathtt{confirm}\,\neg\varphi) \in \mathrm{CS}^+(\mathtt{while}\,\varphi\,\mathtt{do}\,\alpha\,\mathtt{od})$, and $\mathtt{r}^+(i,\gamma)(s) \neq \emptyset$. Since $m \geq k$ this contradicts the definition of k as given above. It follows that $\mathrm{Ter}_{\mathrm{M}}(i, \mathtt{confirm}\,\varphi;\alpha, s') = \mathbf{1}$, which was to be shown.

⊠

4.24. Proposition. *For all* $i \in \mathrm{A}$, $\alpha_1, \alpha_2 \in \mathrm{Ac}^+$ *and* $\varphi \in \mathrm{L}^+$*:*

- $\models^+ \langle \mathrm{do}_i(\alpha_1 + \alpha_2)\rangle\varphi \leftrightarrow (\langle \mathrm{do}_i(\alpha_1)\rangle\varphi \wedge \langle \mathrm{do}_i(\alpha_2)\rangle\varphi)$
- $\models^+ \mathbf{A}_i(\alpha_1 + \alpha_2) \leftrightarrow (\mathbf{A}_i\alpha_1 \wedge \mathbf{A}_i\alpha_2)$

Proof: Let $\mathrm{M} \in \mathbf{M}^+$ with state s, and $\alpha_1, \alpha_2 \in \mathrm{Ac}^+$ and $\varphi \in \mathrm{L}^+$ be arbitrary. We successively show both cases.

$\mathrm{M}, s \models^+ \langle \mathrm{do}_i(\alpha_1 + \alpha_2)\rangle\varphi$
$\Leftrightarrow \forall\alpha' \in \mathrm{CR}^+_{\mathrm{M}}(i, \alpha_1 + \alpha_2, s)\exists s' \in \mathrm{S}(\mathtt{r}^+(i,\alpha')(s) = s' \,\&\, \mathrm{M}, s' \models^+ \varphi)$
$\Leftrightarrow \forall\alpha' \in \mathrm{CR}^+_{\mathrm{M}}(i, \alpha_1, s) \cup \mathrm{CR}^+_{\mathrm{M}}(i, \alpha_2, s)\exists s' \in \mathrm{S}(\mathtt{r}^+(i,\alpha')(s) = s' \,\&\, \mathrm{M}, s' \models^+ \varphi)$
$\Leftrightarrow \forall\alpha' \in \mathrm{CR}^+_{\mathrm{M}}(i, \alpha_1, s)\exists s' \in \mathrm{S}(\mathtt{r}^+(i,\alpha')(s) = s' \,\&\, \mathrm{M}, s' \models^+ \varphi)$ and
$\forall\alpha' \in \mathrm{CR}^+_{\mathrm{M}}(i, \alpha_2, s)\exists s' \in \mathrm{S}(\mathtt{r}^+(i,\alpha')(s) = s' \,\&\, \mathrm{M}, s' \models^+ \varphi)$
$\Leftrightarrow \mathrm{M}, s \models^+ \langle \mathrm{do}_i(\alpha_1)\rangle\varphi$ and $\mathrm{M}, s \models^+ \langle \mathrm{do}_i(\alpha_2)\rangle\varphi$
$\Leftrightarrow \mathrm{M}, s \models^+ \langle \mathrm{do}_i(\alpha_1)\rangle\varphi \wedge \langle \mathrm{do}_i(\alpha_2)\rangle\varphi$

$\mathrm{M}, s \models^+ \mathbf{A}_i(\alpha_1 + \alpha_2)$
$\Leftrightarrow \forall\alpha' \in \mathrm{CR}^+_{\mathrm{M}}(i, \alpha_1 + \alpha_2, s)(\mathtt{c}^+(i,\alpha')(s) = \mathbf{1})$
$\Leftrightarrow \forall\alpha' \in \mathrm{CR}^+_{\mathrm{M}}(i, \alpha_1, s) \cup \mathrm{CR}^+_{\mathrm{M}}(i, \alpha_2, s)(\mathtt{c}^+(i,\alpha')(s) = \mathbf{1})$
$\Leftrightarrow \forall\alpha' \in \mathrm{CR}^+_{\mathrm{M}}(i, \alpha_1, s)(\mathtt{c}^+(i,\alpha')(s) = \mathbf{1})$ and $\forall\alpha' \in \mathrm{CR}^+_{\mathrm{M}}(i, \alpha_2, s)(\mathtt{c}^+(i,\alpha')(s) = \mathbf{1})$
$\Leftrightarrow \mathrm{M}, s \models^+ \mathbf{A}_i\alpha_1$ and $\mathrm{M}, s \models^+ \mathbf{A}_i\alpha_2$
$\Leftrightarrow \mathrm{M}, s \models^+ \mathbf{A}_i\alpha_1 \wedge \mathbf{A}_i\alpha_2$

⊠

4.26. Proposition. *For* $i \in \mathrm{A}$ *we have:*

- $[\mathrm{do}_i(\alpha)](\varphi \to \psi) \to ([\mathrm{do}_i(\alpha)]\varphi \to [\mathrm{do}_i(\alpha)]\psi)$ *is not for all* $\alpha \in \mathrm{Ac}^+$, $\varphi, \psi \in \mathrm{L}^+$, $\models^+$*-valid*
- $\models^+ \psi \Rightarrow \models^+ [\mathrm{do}_i(\alpha)]\psi$ *holds for all* $\alpha \in \mathrm{Ac}^+$, $\psi \in \mathrm{L}^+$

Proof: First note that Definition 4.22 implies that for M a model with state s, and for $i \in \mathrm{A}$, $\alpha \in \mathrm{Ac}^+$ and $\varphi \in \mathrm{L}^+$, it holds that $\mathrm{M}, s \models^+ [\mathrm{do}_i(\alpha)]\varphi$ iff $\exists\alpha' \in \mathrm{CR}^+_{\mathrm{M}}(i, \alpha, s)\forall s' \in \mathrm{S}(\mathtt{r}^+(i,\alpha')(s) = s' \Rightarrow \mathrm{M}, s' \models^+ \varphi)$. The first item is now shown by considering the model $\mathrm{M} = \langle \mathrm{S}, \pi, \mathrm{R}, \mathtt{r}_0, \mathtt{c}_0\rangle$ which is for $p, q \in \Pi$, $i \in \mathrm{A}$ and $a_1, a_2 \in \mathrm{At}$ such that

- $\mathrm{S} = \{s, s_1, s_2\}$
- $\pi(p, s_1) = \mathbf{0}$, $\pi(p, s_2) = \mathbf{1}$, $\pi(q, s_1) = \mathbf{0} = \pi(q, s_2)$
- R is arbitrary

- $\mathtt{r}_0(i,a_1)(s) = s_1$, $\mathtt{r}_0(i,a_2)(s) = s_2$
- $\mathtt{c}_0$ is arbitrary

For this model it holds that

- $\mathrm{M}, s \models^+ [\mathrm{do}_i(a_1 + a_2)](p \to q)$ since $a_1 \in \mathrm{CR}^+_{\mathrm{M}}(i, a_1 + a_2, s)$, $\mathtt{r}^+(i,a_1)(s) = s_1$ and $\mathrm{M}, s_1 \models^+ (p \to q)$
- $\mathrm{M}, s \models^+ [\mathrm{do}_i(a_1 + a_2)]p$ since $a_2 \in \mathrm{CR}^+_{\mathrm{M}}(i, a_1 + a_2, s)$, $\mathtt{r}^+(i,a_2)(s) = s_2$ and $\mathrm{M}, s_2 \models^+ p$
- $\mathrm{M}, s \not\models^+ [\mathrm{do}_i(a_1 + a_2)]q$ since $\mathrm{CR}^+_{\mathrm{M}}(i, a_1 + a_2, s) = \{a_1, a_2\}$, $\mathtt{r}^+(i,a_1)(s) = s_1$ and $\mathtt{r}^+(i,a_2)(s) = s_2$, and both $\mathrm{M}, s_1 \not\models^+ q$ and $\mathrm{M}, s_2 \not\models^+ q$.

These three properties of M suffice to conclude that it is a countermodel to the first item of Proposition 4.26. The second item is a consequence of the fifth item of Proposition 4.23, in which it is stated that $\mathrm{CR}^+_{\mathrm{M}}(i, \alpha, s) \neq \emptyset$ for all $i \in \mathrm{A}$, $\alpha \in \mathrm{Ac}^+$ and states s. For this implies that always some α' exists such that $\alpha' \in \mathrm{CR}^+_{\mathrm{M}}(i, \alpha, s)$, while $\models^+ \psi$ suffices to conclude that for this α' it holds that $\forall s' \in \mathrm{S}(\mathtt{r}^+(i,\alpha')(s) = s' \Rightarrow \mathrm{M}, s' \models^+ \psi)$. ⊠

4.30. Proposition. *Let* M *be a model with state* s *and* U, U_1, U_2 *sets of states in* M. *Then for all* $i \in \mathrm{A}$, $\alpha \in \mathrm{Ac}^+$ *and* s' *in* M *we have:*

1. $(s,U) \in \mathtt{r}^{+,\mathbf{0}}(i,\alpha) \Rightarrow (s' \in U \Leftrightarrow \exists \alpha' \in \mathrm{CS}^+(\alpha)((s,\{s'\}) \in \mathtt{r}^{+,\mathbf{0}}(i,\alpha')))$
2. $(s,U) \in \mathtt{r}^{+,\mathbf{0}}(i,\alpha) \Rightarrow (\mathtt{c}^{+,\mathbf{0}}(i,\alpha)(s) = \mathbf{1} \Leftrightarrow (\mathtt{c}^{+,\mathbf{0}}(i,\alpha')(s) = \mathbf{1}$ *for all* $\alpha' \in \mathrm{CS}^+(\alpha)$ *with* $(s,\{s'\}) \in \mathtt{r}^{+,\mathbf{0}}(i,\alpha')$ *for some* $s' \in U))$
3. $(s,U) \in \mathtt{r}^{+,\mathbf{0}}(i,\alpha) \Rightarrow U \neq \emptyset \,\&\, \exists k \in \mathbb{N}(|U| \leq k)$
4. $(s,U_1) \in \mathtt{r}^{+,\mathbf{0}}(i,\alpha) \,\&\, (s,U_2) \in \mathtt{r}^{+,\mathbf{0}}(i,\alpha) \Rightarrow U_1 = U_2$

where $|U|$ *denotes the cardinality of* U.

Proof: The first and the second item are shown by applying induction on an appropriate ordering on actions. The third item is a slight extension of a proof given by Peleg [101], and the fourth item follows directly from the first one. Here we sketch the proof of the first item. To show this item we prove the following two implications:

'$\Rightarrow$' $((s,U) \in \mathtt{r}^{+,\mathbf{0}}(i,\alpha) \,\&\, s' \in U) \Rightarrow \exists \alpha' \in \mathrm{CS}^+(\alpha)((s,\{s'\}) \in \mathtt{r}^{+,\mathbf{0}}(i,\alpha'))$

'$\Leftarrow$' $((s,U) \in \mathtt{r}^{+,\mathbf{0}}(i,\alpha) \,\&\, \exists \alpha' \in \mathrm{CS}^+(\alpha)((s,\{s'\}) \in \mathtt{r}^{+,\mathbf{0}}(i,\alpha'))) \Rightarrow s' \in U$

The '$\Rightarrow$'-case is shown by induction on the subaction ordering $<$ which is defined as the transitive closure of $\prec$, where $\prec$ is the smallest relation on $\mathrm{Ac}^+ \times \mathrm{Ac}^+$ that satisfies for all $\varphi \in \mathrm{L}^+$ and $\alpha, \alpha_1, \alpha_2 \in \mathrm{Ac}^+$ the following constraints:

1. $\alpha_1 \prec \alpha_1; \alpha_2$
2. $\alpha_2 \prec \alpha_1; \alpha_2$
3. $\mathtt{confirm}\,\varphi; \alpha_1 \prec \mathtt{if}\,\varphi\,\mathtt{then}\,\alpha_1\,\mathtt{else}\,\alpha_2\,\mathtt{fi}$
4. $\mathtt{confirm}\,\neg\varphi; \alpha_2 \prec \mathtt{if}\,\varphi\,\mathtt{then}\,\alpha_1\,\mathtt{else}\,\alpha_2\,\mathtt{fi}$
5. $\mathtt{confirm}\,\varphi; \alpha \prec \mathtt{while}\,\varphi\,\mathtt{do}\,\alpha\,\mathtt{od}$

6. $\alpha_1 \prec \alpha_1 + \alpha_2$
7. $\alpha_2 \prec \alpha_1 + \alpha_2$

The well-foundedness of $<$ is shown using the lexicographic path ordering, the technique which was also applied in the soundness and completeness proofs of Chapter 3. We show '$\Rightarrow$' for the case that α is a repetitive composition; the other cases are more easy and left to the reader.

To show the case for repetitive compositions, let $i \in \mathrm{A}$, $\varphi \in \mathrm{L}^+$ and $\alpha \in \mathrm{Ac}^+$ be such that for some model M with state s it holds that $(s,U) \in \mathtt{r}^{+,\mathbf{0}}(i, \mathtt{while}\,\varphi\,\mathtt{do}\,\alpha\,\mathtt{od})$ while $s' \in U$. Assume that F_k is defined as on page 87. Then it holds that $(s,U) \in \mathrm{F}_l$ for some l. We show by induction that for all k holds that whenever $(s,U) \in \mathrm{F}_k$ and $s' \in U$ then $(s,\{s'\}) \in \mathtt{r}^{+,\mathbf{0}}(i,\alpha')$ for some $\alpha' \in \mathrm{CS}^+(\mathtt{while}\,\varphi\,\mathtt{do}\,\alpha\,\mathtt{od})$.

$k = 0$: Since $\mathrm{F}_0 = \mathtt{r}^{+,\mathbf{0}}(i, \mathtt{confirm}\,\neg\varphi)$, it follows that $(s,U) \in \mathrm{F}_0$ implies $U = \{s\}$ and $\mathrm{M}, s \models^{+,\mathbf{0}} \varphi$. Then it holds that $(s,\{s\}) \in \mathtt{r}^{+,\mathbf{0}}(i, \mathtt{confirm}\,\neg\varphi)$ while $\mathtt{confirm}\,\neg\varphi \in \mathrm{CS}^+(\mathtt{while}\,\varphi\,\mathtt{do}\,\alpha\,\mathtt{od})$, which suffices to conclude the case for $k = 0$.

$k \mapsto k+1$: If $(s,U) \in \mathrm{F}_{k+1}$, then either $(s,U) \in \mathrm{F}_0$ or $(s,U) \in \mathtt{r}^{+,\mathbf{0}}(i, \mathtt{confirm}\,\varphi; \alpha) \cdot \mathrm{F}_k$. If $(s,U) \in \mathrm{F}_0$ then the case is shown as above. So assume that $(s,U) \in \mathtt{r}^{+,\mathbf{0}}(i, \mathtt{confirm}\,\varphi; \alpha) \cdot \mathrm{F}_k$. Then by definition of $\cdot$, $s_1, U_1, s_2, U_2, \ldots$ exist such that $(s,\{s_1, s_2, \ldots\}) \in \mathtt{r}^{+,\mathbf{0}}(i, \mathtt{confirm}\,\varphi; \alpha)$ and $(s_m, U_m) \in \mathrm{F}_k$ while $U = \bigcup_m U_m$. By applying the induction hypothesis for k we have that some $\beta \in \mathrm{CS}^+(\mathtt{while}\,\varphi\,\mathtt{do}\,\alpha\,\mathtt{od})$ exists such that $(s_m, t_m) \in \mathtt{r}^{+,\mathbf{0}}(i,\beta)$, for all $t_m \in U_m$. By induction on $<$ some $\gamma \in \mathrm{CS}^+(\mathtt{confirm}\,\varphi; \alpha)$ exists such that $(s, s_m) \in \mathtt{r}^{+,\mathbf{0}}(i,\gamma)$. Then by definition of CS^+ we have that $\beta;\gamma \in \mathrm{CS}^+(\mathtt{while}\,\varphi\,\mathtt{do}\,\alpha\,\mathtt{od})$, and since $(s,t_m) \in \mathtt{r}^{+,\mathbf{0}}(i,\beta;\gamma)$ the statement follows for $k+1$.

k is a limit-ordinal: This case follows directly by induction hypothesis.

Since the case holds for all k, it in particular holds for F_l, which, since (s,U) was assumed to be in F_l, suffices to conclude '$\Rightarrow$'.

The case for '$\Leftarrow$' is also shown by induction on the subaction ordering. Again we sketch this case for repetitive compositions. Assume that $(s,U) \in \mathtt{r}^{+,\mathbf{0}}(i, \mathtt{while}\,\varphi\,\mathtt{do}\,\alpha\,\mathtt{od})$ and that for some $\alpha' \in \mathrm{CS}^+(\mathtt{while}\,\varphi\,\mathtt{do}\,\alpha\,\mathtt{od})$ holds that $(s,\{s'\}) \in \mathtt{r}^{+,\mathbf{0}}(i,\alpha')$. We have to show that $s' \in U$. Now if $\mathrm{M}, s \models^{+,\mathbf{0}} \neg\varphi$ then $\alpha' = \mathtt{confirm}\,\neg\varphi$, $s' = s$, and $U = \{s\}$, which indeed implies that $s' \in U$. So assume that $\mathrm{M}, s \models^{+,\mathbf{0}} \varphi$. It is obvious that in this case $(\mathtt{confirm}\,\varphi;\beta)$, with $\beta \in \mathrm{CS}^+(\alpha)$, is a prefix of α'. It is even so obvious that $(s,U) \in \mathtt{r}^{+,\mathbf{0}}(i, \mathtt{confirm}\,\varphi; \alpha) \cdot \mathrm{F}_l$ for some l. This implies that states $u_1, u_2, \ldots$ and sets $U_1, U_2, \ldots$ of states exist such that $(s,\{u_1, u_2, \ldots\}) \in \mathtt{r}^{+,\mathbf{0}}(i,\alpha)$, $(u_m, U_m) \in \mathrm{F}_l$, and $U = \bigcup_m U_m$. By the induction hypothesis it follows that s_1 such that $(s,\{s_1\}) \in \mathtt{r}^{+,\mathbf{0}}(i,\beta)$ is a member of the set $\{u_1, u_2, \ldots\}$, say $s_1 = u_k$. This leaves us in the situation where $(u_k, U_k) \in \mathtt{r}^{+,\mathbf{0}}(i, \mathtt{while}\,\varphi\,\mathtt{do}\,\alpha\,\mathtt{od})$, and $(u_k,\{s'\}) \in \mathtt{r}^{+,\mathbf{0}}(i,\beta')$ for β' such that $\alpha' = (\mathtt{confirm}\,\varphi;\beta);\beta'$. Now if $\mathrm{M}, u_k \models^{+,\mathbf{0}} \neg\varphi$ then $s' = u_k$ and we are done. Otherwise we proceed analogously

until we reach the state s'; note that, since $\mathrm{M}, s' \models^{+,\mathbf{0}} \neg\varphi$, this process terminates.

From the proofs of the cases '$\Rightarrow$' and '$\Leftarrow$' we conclude that the claim stated in item 1 of Proposition 4.30 holds for repetitive compositions.
⊠

4.31. PROPOSITION. *For all* $i \in \mathrm{A}$, $\alpha, \alpha_1, \alpha_2 \in \mathrm{Ac}^+$ *and* $\varphi, \psi \in \mathrm{L}^+$ *we have:*

1. $\models^{+,\mathbf{0}} \langle \mathrm{do}_i(\mathtt{confirm}\,\varphi)\rangle\psi \leftrightarrow (\varphi \wedge \psi)$
2. $\models^{+,\mathbf{0}} \langle \mathrm{do}_i(\alpha_1;\alpha_2)\rangle\psi \leftrightarrow \langle \mathrm{do}_i(\alpha_1)\rangle\langle \mathrm{do}_i(\alpha_2)\rangle\psi$
3. $\models^{+,\mathbf{0}} \langle \mathrm{do}_i(\mathtt{if}\,\varphi\,\mathtt{then}\,\alpha_1\,\mathtt{else}\,\alpha_2\,\mathtt{fi})\rangle\psi \leftrightarrow ((\varphi \wedge \langle \mathrm{do}_i(\alpha_1)\rangle\psi) \vee (\neg\varphi \wedge \langle \mathrm{do}_i(\alpha_2)\rangle\psi))$
4. $\models^{+,\mathbf{0}} \langle \mathrm{do}_i(\mathtt{while}\,\varphi\,\mathtt{do}\,\alpha\,\mathtt{od})\rangle\psi \leftrightarrow ((\neg\varphi \wedge \psi) \vee (\varphi \wedge \langle \mathrm{do}_i(\alpha)\rangle\langle \mathrm{do}_i(\mathtt{while}\,\varphi\,\mathtt{do}\,\alpha\,\mathtt{od})\rangle\psi))$
5. $\models^{+,\mathbf{0}} \langle \mathrm{do}_i(\alpha_1 + \alpha_2)\rangle\varphi \leftrightarrow (\langle \mathrm{do}_i(\alpha_1)\rangle\varphi \wedge \langle \mathrm{do}_i(\alpha_2)\rangle\varphi)$

PROOF: We show the first, second, fourth and fifth item, leaving the third item to the reader. Let $\mathrm{M} \in \mathbf{M}^+$ with state s, and $i \in \mathrm{A}$, $\alpha, \alpha_1, \alpha_2 \in \mathrm{Ac}^+$ and $\varphi, \psi \in \mathrm{L}^+$ be arbitrary.

1. $\mathrm{M}, s \models^+ \langle \mathrm{do}_i(\mathtt{confirm}\,\varphi)\rangle\psi$
   $\Leftrightarrow \exists U((s,U) \in \mathtt{r}^{+,\mathbf{0}}(i, \mathtt{confirm}\,\varphi)\,\&\,\mathrm{M}, s' \models^{+,\mathbf{0}} \psi$ for all $s' \in U)$
   $\Leftrightarrow (s,\{s\}) \in \mathtt{r}^{+,\mathbf{0}}(i, \mathtt{confirm}\,\varphi)\,\&\,\mathrm{M}, s' \models^{+,\mathbf{0}} \psi$ for all $s' \in \{s\}$
   $\Leftrightarrow \mathrm{M}, s \models^{+,\mathbf{0}} \varphi\,\&\,\mathrm{M}, s \models^{+,\mathbf{0}} \psi$
   $\Leftrightarrow \mathrm{M}, s \models^{+,\mathbf{0}} \varphi \wedge \psi$
2. $\mathrm{M}, s \models^{+,\mathbf{0}} \langle \mathrm{do}_i(\alpha_1;\alpha_2)\rangle\psi$
   $\Leftrightarrow \exists U((s,U) \in \mathtt{r}^{+,\mathbf{0}}(i, \alpha_1;\alpha_2)\,\&\,\mathrm{M}, s' \models^{+,\mathbf{0}} \psi$ for all $s' \in U)$
   $\Leftrightarrow \exists U((s,U) \in \mathtt{r}^{+,\mathbf{0}}(i, \alpha_1) \cdot \mathtt{r}^{+,\mathbf{0}}(i, \alpha_2)\,\&\,\mathrm{M}, s' \models^{+,\mathbf{0}} \psi$ for all $s' \in U)$
   $\Leftrightarrow \exists s_1, U_1 \exists s_2, U_2 \ldots((s,\{s_1, s_2, \ldots\}) \in \mathtt{r}^{+,\mathbf{0}}(i, \alpha_1)\,\&$
   $\forall k((s_k, U_k) \in \mathtt{r}^{+,\mathbf{0}}(i, \alpha_2)\,\&\,\mathrm{M}, s' \models^{+,\mathbf{0}} \psi$ for all $s' \in U_k))$
   $\Leftrightarrow \exists s_1 \exists s_2 \ldots((s,\{s_1, s_2, \ldots\}) \in \mathtt{r}^{+,\mathbf{0}}(i, \alpha_1)\,\&\,\forall k(\mathrm{M}, s_k \models^{+,\mathbf{0}} \langle \mathrm{do}_i(\alpha_2)\rangle\psi))$
   $\Leftrightarrow \mathrm{M}, s \models^{+,\mathbf{0}} \langle \mathrm{do}_i(\alpha_1)\rangle\langle \mathrm{do}_i(\alpha_2)\rangle\psi$
4. $\mathrm{M}, s \models^{+,\mathbf{0}} \langle \mathrm{do}_i(\mathtt{while}\,\varphi\,\mathtt{do}\,\alpha\,\mathtt{od})\rangle\psi$
   $\Leftrightarrow \exists U((s,U) \in \mathtt{r}^{+,\mathbf{0}}(i, \mathtt{while}\,\varphi\,\mathtt{do}\,\alpha\,\mathtt{od})\,\&\,\mathrm{M}, s' \models^{+,\mathbf{0}} \psi$ for all $s' \in U)$
   $\Leftrightarrow \exists U((s,U) \in \mathrm{LFP}(\mathrm{F}_{(i,\mathtt{confirm}\,\varphi;\alpha)})\,\&\,\mathrm{M}, s' \models^{+,\mathbf{0}} \psi$ for all $s' \in U)$
   $\Leftrightarrow \exists U((s,U) \in \mathrm{F}_{(i,\mathtt{confirm}\,\varphi;\alpha)}(\mathrm{LFP}(\mathrm{F}_{(i,\mathtt{confirm}\,\varphi;\alpha)}))\,\&$
   $\mathrm{M}, s' \models^{+,\mathbf{0}} \psi$ for all $s' \in U)$
   $\Leftrightarrow \exists U((s,U) \in \mathtt{r}^{+,\mathbf{0}}(i, \mathtt{confirm}\,\neg\varphi) \cup \mathtt{r}^{+,\mathbf{0}}(i, \mathtt{confirm}\,\varphi;\alpha) \cdot \mathrm{LFP}(\mathrm{F}_{(i,\mathtt{confirm}\,\varphi;\alpha)})\,\&$
   $\mathrm{M}, s' \models^{+,\mathbf{0}} \psi$ for all $s' \in U)$
   $\Leftrightarrow \exists U((s,U) \in \mathtt{r}^{+,\mathbf{0}}(i, \mathtt{confirm}\,\neg\varphi)\,\&\,\mathrm{M}, s' \models^{+,\mathbf{0}} \psi$ for all $s' \in U)$ or
   $\exists U((s,U) \in \mathtt{r}^{+,\mathbf{0}}(i, \mathtt{confirm}\,\varphi;\alpha) \cdot \mathrm{LFP}(\mathrm{F}_{(i,\mathtt{confirm}\,\varphi;\alpha)})\,\&$
   $\mathrm{M}, s' \models^{+,\mathbf{0}} \psi$ for all $s' \in U)$
   $\Leftrightarrow \mathrm{M}, s \models^{+,\mathbf{0}} \langle \mathrm{do}_i(\mathtt{confirm}\,\neg\varphi)\rangle\psi$ or
   $\exists U((s,U) \in \mathtt{r}^{+,\mathbf{0}}(i, \mathtt{confirm}\,\varphi;\alpha) \cdot \mathtt{r}^{+,\mathbf{0}}(i, \mathtt{while}\,\varphi\,\mathtt{do}\,\alpha\,\mathtt{od})\,\&$
   $\mathrm{M}, s' \models^{+,\mathbf{0}} \psi$ for all $s' \in U)$

$\Leftrightarrow$ $\mathrm{M}, s \models^{+,\mathbf{0}} (\neg\varphi \wedge \psi)$ or
$\exists U((s,U) \in \mathtt{r}^{+,\mathbf{0}}(i, (\mathtt{confirm}\,\varphi; \alpha); \mathtt{while}\,\varphi\,\mathtt{do}\,\alpha\,\mathtt{od}) \,\&\, \mathrm{M}, s' \models^{+,\mathbf{0}} \psi$ for all $s' \in U)$
$\Leftrightarrow$ $\mathrm{M}, s \models^{+,\mathbf{0}} (\neg\varphi \wedge \psi)$ or $\mathrm{M}, s \models^{+,\mathbf{0}} \langle \mathrm{do}_i((\mathtt{confirm}\,\varphi; \alpha); \mathtt{while}\,\varphi\,\mathtt{do}\,\alpha\,\mathtt{od})\rangle\psi$
$\Leftrightarrow$ $\mathrm{M}, s \models^{+,\mathbf{0}} (\neg\varphi \wedge \psi) \vee (\varphi \wedge \langle \mathrm{do}_i(\alpha)\rangle\langle \mathrm{do}_i(\mathtt{while}\,\varphi\,\mathtt{do}\,\alpha\,\mathtt{od})\rangle\psi)$

The last equivalence but one comprises three steps: in the first two of these, $\langle \mathrm{do}_i((\mathtt{confirm}\,\varphi; \alpha); \mathtt{while}\,\varphi\,\mathtt{do}\,\alpha\,\mathtt{od})\rangle\psi$ is, by twice applying item 2, transformed into $\langle \mathrm{do}_i(\mathtt{confirm}\,\varphi)\rangle\langle \mathrm{do}_i(\alpha)\rangle\langle \mathrm{do}_i(\mathtt{while}\,\varphi\,\mathtt{do}\,\alpha\,\mathtt{od})\rangle\psi$, which, using the first item, is transformed to $\varphi \wedge \langle \mathrm{do}_i(\alpha)\rangle\langle \mathrm{do}_i(\mathtt{while}\,\varphi\,\mathtt{do}\,\alpha\,\mathtt{od})\rangle\psi$.

5. $\mathrm{M}, s \models^{+,\mathbf{0}} \langle \mathrm{do}_i(\alpha_1 + \alpha_2)\rangle\varphi$
$\Leftrightarrow$ $\exists U((s,U) \in \mathtt{r}^{+,\mathbf{0}}(i, \alpha_1 + \alpha_2) \,\&\, \mathrm{M}, s' \models^{+,\mathbf{0}} \varphi$ for all $s' \in U)$
$\Leftrightarrow$ $\exists U_1 \exists U_2((s,U_1) \in \mathtt{r}^{+,\mathbf{0}}(i, \alpha_1) \,\&\, (s,U_2) \in \mathtt{r}^{+,\mathbf{0}}(i, \alpha_2) \,\&$
$\mathrm{M}, s' \models^{+,\mathbf{0}} \psi$ for all $s' \in U_1 \cup U_2)$
$\Leftrightarrow$ $\exists U_1((s,U_1) \in \mathtt{r}^{+,\mathbf{0}}(i, \alpha_1) \,\&\, \mathrm{M}, s' \models^{+,\mathbf{0}} \psi$ for all $s' \in U_1)$ and
$\exists U_2((s,U_2) \in \mathtt{r}^{+,\mathbf{0}}(i, \alpha_2) \,\&\, \mathrm{M}, s' \models^{+,\mathbf{0}} \psi$ for all $s' \in U_2)$
$\Leftrightarrow$ $\mathrm{M}, s \models^{+,\mathbf{0}} \langle \mathrm{do}_i(\alpha_1)\rangle\psi$ and $\mathrm{M}, s \models^{+,\mathbf{0}} \langle \mathrm{do}_i(\alpha_2)\rangle\psi$
$\Leftrightarrow$ $\mathrm{M}, s \models^{+,\mathbf{0}} \langle \mathrm{do}_i(\alpha_1)\rangle\psi \wedge \langle \mathrm{do}_i(\alpha_2)\psi$

4.32. Proposition. *For $i \in \mathrm{A}$ we have:*

- $[\mathrm{do}_i(\alpha)](\varphi \to \psi) \to ([\mathrm{do}_i(\alpha)]\varphi \to [\mathrm{do}_i(\alpha)]\psi)$ *is not for all $\alpha \in \mathrm{Ac}^+$, $\varphi, \psi \in \mathrm{L}^+$, $\models^{+,\mathbf{0}}$-valid*
- $\models^{+,\mathbf{0}} \psi \Rightarrow \models^{+,\mathbf{0}} [\mathrm{do}_i(\alpha)]\psi$ *holds for all $\alpha \in \mathrm{Ac}^+$, $\psi \in \mathrm{L}^+$*

Proof: First note that Definition 4.28 implies that for M a model with state s, and for $i \in \mathrm{A}$, $\alpha \in \mathrm{Ac}^+$ and $\varphi \in \mathrm{L}^+$, it holds that $\mathrm{M}, s \models^{+,\mathbf{0}} [\mathrm{do}_i(\alpha)]\varphi$ iff $\forall U((s,U) \in \mathtt{r}^{+,\mathbf{0}}(i,\alpha) \Rightarrow \exists s' \in U(\mathrm{M}, s' \models^{+,\mathbf{0}} \varphi))$. The first item is now shown by the same model used to show the first item of Proposition 4.26. The second item follows from the third item of Proposition 4.30. For then all the sets U such that $(s,U) \in \mathtt{r}^{+,\mathbf{0}}(i,\alpha)$ are nonempty, while $\models^{+,\mathbf{0}} \psi$ guarantees that all the states $s' \in U$ satisfy ψ. This suffices to conclude that $\models^{+,\mathbf{0}} [\mathrm{do}_i(\alpha)]\psi$ holds.
⊠

4.33. Proposition. *For all $i \in \mathrm{A}$, $\alpha, \alpha_1, \alpha_2 \in \mathrm{Ac}^+$ and $\varphi \in \mathrm{L}^+$ we have:*

1. $\models^{+,\mathbf{0}} \mathbf{A}_i\mathtt{confirm}\,\varphi \leftrightarrow \varphi$
2. $\models^{+,\mathbf{0}} \mathbf{A}_i\alpha_1; \alpha_2 \leftrightarrow \mathbf{A}_i\alpha_1 \wedge \langle \mathrm{do}_i(\alpha_1)\rangle\mathbf{A}_i\alpha_2$
3. $\models^{+,\mathbf{0}} \mathbf{A}_i\mathtt{if}\,\varphi\,\mathtt{then}\,\alpha_1\,\mathtt{else}\,\alpha_2\,\mathtt{fi} \leftrightarrow ((\varphi \wedge \mathbf{A}_i\alpha_1) \vee (\neg\varphi \wedge \mathbf{A}_i\alpha_2))$
4. $\models^{+,\mathbf{0}} \mathbf{A}_i\mathtt{while}\,\varphi\,\mathtt{do}\,\alpha\,\mathtt{od} \leftrightarrow (\neg\varphi \vee (\varphi \wedge \mathbf{A}_i\alpha \wedge \langle \mathrm{do}_i(\alpha)\rangle\mathbf{A}_i\mathtt{while}\,\varphi\,\mathtt{do}\,\alpha\,\mathtt{od}))$
5. $\models^{+,\mathbf{0}} \mathbf{A}_i\alpha_1 + \alpha_2 \leftrightarrow (\mathbf{A}_i\alpha_1 \wedge \mathbf{A}_i\alpha_2)$

Proof: We show the second and fourth item; the other items are similar to previously proved ones. Let $\mathrm{M} \in \mathbf{M}^+$ with state s and $\alpha, \alpha_1, \alpha_2 \in \mathrm{Ac}^+$ and $\varphi \in \mathrm{L}^+$ be arbitrary.

$$
\begin{aligned}
&\mathrm{M}, s \models^{+,\mathbf{0}} \mathbf{A}_i\alpha_1;\alpha_2\\
\Leftrightarrow\ &\mathtt{c}^{+,\mathbf{0}}(i,\alpha_1;\alpha_2)(s) = \mathbf{1}\\
\Leftrightarrow\ &\mathtt{c}^{+,\mathbf{0}}(i,\alpha_1)(s) = \mathbf{1} \,\&\, \exists U((s,U) \in \mathtt{r}^{+,\mathbf{0}}(i,\alpha_1) \,\&\, \mathtt{c}^{+,\mathbf{0}}(i,\alpha_2)(s') = \mathbf{1} \text{ for all } s' \in U)\\
\Leftrightarrow\ &\mathtt{c}^{+,\mathbf{0}}(i,\alpha_1)(s) = \mathbf{1} \,\&\, \exists U((s,U) \in \mathtt{r}^{+,\mathbf{0}}(i,\alpha_1) \,\&\, \mathrm{M}, s' \models^{+,\mathbf{0}} \mathbf{A}_i\alpha_2 \text{ for all } s' \in U)\\
\Leftrightarrow\ &\mathrm{M}, s \models^{+,\mathbf{0}} \mathbf{A}_i\alpha_1 \,\&\, \mathrm{M}, s \models^{+,\mathbf{0}} \langle \mathrm{do}_i(\alpha_1)\rangle \mathbf{A}_i\alpha_2\\
\Leftrightarrow\ &\mathrm{M}, s \models^{+,\mathbf{0}} \mathbf{A}_i\alpha_1 \wedge \langle \mathrm{do}_i(\alpha_1)\rangle \mathbf{A}_i\alpha_2
\end{aligned}
$$

To show the fourth item we prove that if $\mathrm{M}, s \models^{+,\mathbf{0}} \varphi$ then $\mathrm{M}, s \models^{+,\mathbf{0}} \mathbf{A}_i\mathtt{while}\,\varphi\,\mathtt{do}\,\alpha\,\mathtt{od}$ iff $\mathrm{M}, s \models^{+,\mathbf{0}} \mathbf{A}_i(\mathtt{confirm}\,\varphi;\alpha);\mathtt{while}\,\varphi\,\mathtt{do}\,\alpha\,\mathtt{od}$, which, using items 1 and 2, suffices to conclude that item 4 indeed holds. To this end we use the following lemma, the proof of which is straightforward by definition of CS^+ and $\mathtt{r}^{+,\mathbf{0}}$.

4.36. Lemma. *Let* M *be a model with state* s *and let* U *be a set of states in* M. *For all* $i \in \mathrm{A}$, $\varphi \in \mathrm{L}^+$ *and* $\alpha \in \mathrm{Ac}^+$ *we have:*

$$
\begin{aligned}
&\mathrm{M}, s \models^{+,\mathbf{0}} \varphi \Rightarrow\\
&((s,U) \in \mathtt{r}^{+,\mathbf{0}}(i,\mathtt{while}\,\varphi\,\mathtt{do}\,\alpha\,\mathtt{od}) \Leftrightarrow (s,U) \in \mathtt{r}^{+,\mathbf{0}}(i,(\mathtt{confirm}\,\varphi;\alpha);\mathtt{while}\,\varphi\,\mathtt{do}\,\alpha\,\mathtt{od}))
\end{aligned}
$$

Now let M with state s be such that $\mathrm{M}, s \models^{+,\mathbf{0}} \varphi$. We show the equivalence of $\mathrm{M}, s \models^{+,\mathbf{0}} \mathbf{A}_i\mathtt{while}\,\varphi\,\mathtt{do}\,\alpha\,\mathtt{od}$ and $\mathrm{M}, s \models^{+,\mathbf{0}} \mathbf{A}_i(\mathtt{confirm}\,\varphi;\alpha);\mathtt{while}\,\varphi\,\mathtt{do}\,\alpha\,\mathtt{od}$ by proving two implications.

'$\Rightarrow$' Assume that $\mathrm{M}, s \models^{+,\mathbf{0}} \mathbf{A}_i\mathtt{while}\,\varphi\,\mathtt{do}\,\alpha\,\mathtt{od}$, i.e. $\mathtt{c}^{+,\mathbf{0}}(i,\mathtt{while}\,\varphi\,\mathtt{do}\,\alpha\,\mathtt{od})(s) = \mathbf{1}$. This implies that for some set U of states in M, $(s,U) \in \mathtt{r}^{+,\mathbf{0}}(i,\mathtt{while}\,\varphi\,\mathtt{do}\,\alpha\,\mathtt{od})$ and $\forall s' \in U \forall \beta \in \mathrm{CS}^+(\mathtt{while}\,\varphi\,\mathtt{do}\,\alpha\,\mathtt{od})((s,\{s'\}) \in \mathtt{r}^{+,\mathbf{0}}(i,\beta) \Rightarrow \mathtt{c}^{+,\mathbf{0}}(i,\beta)(s) = \mathbf{1})$. Now since $\mathrm{M}, s \models^{+,\mathbf{0}} \varphi$ it follows that $(s,U) \in \mathtt{r}^{+,\mathbf{0}}(i,(\mathtt{confirm}\,\varphi;\alpha);\mathtt{while}\,\varphi\,\mathtt{do}\,\alpha\,\mathtt{od})$ by Lemma 4.36. Let $s' \in U$ and assume that $\gamma \in \mathrm{CS}^+((\mathtt{confirm}\,\varphi;\alpha);\mathtt{while}\,\varphi\,\mathtt{do}\,\alpha\,\mathtt{od})$ is such that $(s,\{s'\}) \in \mathtt{r}^{+,\mathbf{0}}(\gamma)$; such a γ exists by Proposition 4.30(1). Now, by definition of CS^+, we have $\gamma \in \mathrm{CS}^+(\mathtt{while}\,\varphi\,\mathtt{do}\,\alpha\,\mathtt{od})$. Since $\mathtt{c}^{+,\mathbf{0}}(i,\mathtt{while}\,\varphi\,\mathtt{do}\,\alpha\,\mathtt{od})(s) = \mathbf{1}$ it follows by the left-to-right implication of the second item of Proposition 4.30 that $\mathtt{c}^{+,\mathbf{0}}(i,\gamma)(s) = \mathbf{1}$. By the right-to-left implication of the same item it follows that $\mathtt{c}^{+,\mathbf{0}}(i,(\mathtt{confirm}\,\varphi;\alpha);\mathtt{while}\,\varphi\,\mathtt{do}\,\alpha\,\mathtt{od})(s) = \mathbf{1}$, which is necessary and sufficient to conclude that $\mathrm{M}, s \models^{+,\mathbf{0}} \mathbf{A}_i(\mathtt{confirm}\,\varphi;\alpha);\mathtt{while}\,\varphi\,\mathtt{do}\,\alpha\,\mathtt{od}$.

'$\Leftarrow$' Assume that $\mathrm{M}, s \models^{+,\mathbf{0}} \mathbf{A}_i(\mathtt{confirm}\,\varphi;\alpha);\mathtt{while}\,\varphi\,\mathtt{do}\,\alpha\,\mathtt{od}$. From this it follows that $\mathtt{c}^{+,\mathbf{0}}(i,\mathtt{confirm}\,\varphi;\alpha)(s) = \mathbf{1}$ and some set U of states exists such that $(s,U) \in \mathtt{r}^{+,\mathbf{0}}(i,\mathtt{confirm}\,\varphi;\alpha)$ and $\mathtt{c}^{+,\mathbf{0}}(i,\mathtt{while}\,\varphi\,\mathtt{do}\,\alpha\,\mathtt{od})(s') = \mathbf{1}$ for all $s' \in U$. Assume that $U = \{s_1, s_2, \ldots\}$. Then $\mathtt{c}^{+,\mathbf{0}}(i,\mathtt{while}\,\varphi\,\mathtt{do}\,\alpha\,\mathtt{od})(s_k) = \mathbf{1}$ which implies that some set U_k exists with $(s_k,U_k) \in \mathtt{r}^{+,\mathbf{0}}(i,\mathtt{while}\,\varphi\,\mathtt{do}\,\alpha\,\mathtt{od})$. Now define $V \triangleq \bigcup_k U_k$. It follows that $(s,V) \in \mathtt{r}^{+,\mathbf{0}}(i,\mathtt{confirm}\,\varphi;\alpha) \cdot \mathtt{r}^{+,\mathbf{0}}(i,\mathtt{while}\,\varphi\,\mathtt{do}\,\alpha\,\mathtt{od})$, which, by the fact that $\mathtt{r}^{+,\mathbf{0}}(i,\mathtt{while}\,\varphi\,\mathtt{do}\,\alpha\,\mathtt{od})$ is a fixed point of $\mathrm{F}_{(i,\mathtt{confirm}\varphi;\alpha)}$, implies that $(s,V) \in \mathtt{r}^{+,\mathbf{0}}(i,\mathtt{while}\,\varphi\,\mathtt{do}\,\alpha\,\mathtt{od})$. Now let $s' \in V$. By the first item of Proposition 4.30 it follows that $(s,\{s'\}) \in \mathtt{r}^{+,\mathbf{0}}(i,\beta)$ for some $\beta \in \mathrm{CS}^+(\mathtt{while}\,\varphi\,\mathtt{do}\,\alpha\,\mathtt{od})$. Since $\mathrm{M}, s \models^{+,\mathbf{0}}$

φ it follows that $\beta \in \mathrm{CS}^{+}((\texttt{confirm}\,\varphi;\alpha);\texttt{while}\,\varphi\,\texttt{do}\,\alpha\,\texttt{od})$. From the fact that $\mathsf{c}^{+,\mathbf{0}}(i,(\texttt{confirm}\,\varphi;\alpha);\texttt{while}\,\varphi\,\texttt{do}\,\alpha\,\texttt{od})(s) = \mathbf{1}$ it follows by the left-to-right implication of the second item of Proposition 4.30 that $\mathsf{c}^{+,\mathbf{0}}(i,\beta)(s) = \mathbf{1}$. By the right-to-left implication of the second item of Proposition 4.30, $\mathsf{c}^{+,\mathbf{0}}(i,\texttt{while}\,\varphi\,\texttt{do}\,\alpha\,\texttt{od})(s) = \mathbf{1}$, and hence $\mathrm{M}, s \models^{+,\mathbf{0}} \mathbf{A}_i\texttt{while}\,\varphi\,\texttt{do}\,\alpha\,\texttt{od}$, which was to be shown.

⊠

Chapter 5

Intelligent information agents

Is this a dagger which I see before me,
The handle toward my hand?
Come, let me clutch thee.
I have thee not, and yet I see thee still.
Art thou not, fatal vision, sensible
To feeling as to sight? or art thou but
A dagger of the mind, a false creation,
Proceeding from the heat-oppressed brain?

Shakespeare, *'Macbeth, Act 2, Scene 1'*.

In this chapter we present a framework that allows one to formalise intelligent information agents, which are agents that manipulate information, usually on the authority of some user they assist with his/her information management. The formalisation that we present concerns the agents' informational attitudes, the actions that they may perform and the interaction between these notions. We consider various notions of belief that differ in the degree of credibility that is attached to them. In combination with knowledge, which is treated as the most credible kind of belief, these notions constitute the agents' informational attitudes. In addition we formalise special, so-called informative actions. By performing an informative action an agent may affect a designated region of its beliefs, i.e. each of the informative actions is associated with a special part of the agent's beliefs. The informative actions that we propose can be seen as corresponding to three different ways in which an agent can acquire information, viz. through observations, through communication, and by making assumptions by default. Of these, observations are considered to yield information with the highest credibility, whereas information adopted by default is assumed to have the lowest degree of credibility. In general, whenever an agent successfully executes an informative action all regions of the same, or, in the case of communication, another agent's beliefs are modified of which the credibility is at most that of the information acquired as the result of executing the informative action. To

account for this modification of an agent's beliefs we interpret informative actions as model-transformers, rather than the more common interpretation of actions as state-transitions that was employed in the previous chapters. This interpretation is such that the changes in the beliefs of an agent brought about by execution of an informative action comply with the well-known AGM postulates for belief revision. Abilities for informative actions are used to formalise both the limited capacities of agents to acquire information and the preference of one kind of information acquisition over another. As usual, we conclude this chapter with a short summary, an indication of possible extensions, some pointers to the relevant literature, and a collection of proofs of selected propositions.

5.1 Intelligent information agents

An important class of agents' implementations deals with agents that manipulate information. These intelligent information agents as we call them are usually software agents that assist a (computer) user with his/her information management, thereby helping the user to sift his/her way through the information age. Agents like these may for instance manage the user's email by automatically sorting or forwarding incoming email messages. Also implemented agents exist that manage the Usenet newsgroups for a user, for instance by filtering out messages that it considers (ir)relevant for the user. The essence of these intelligent information agents is the interaction between their acts and their information: on the one hand these agents perform actions based on the information they have, on the other hand they may perform actions to acquire additional information.

The agents that we formalise in this chapter may be seen as (moderately) intelligent information agents that have three possible sources of information at their disposal. The first of these is an *exogenous* source of information and consists of observations that an agent makes about the current world. The software agents implemented at the MIT Media Laboratory [88, 89, 90] use this source of information to learn about the user they are assisting. The second source is also an exogenous one and is constituted by the information that other agents communicate. For example, the intelligent information agent that assists some user with his/her information management on the World Wide Web may contact an agent at another site to find out about the availability of technical reports that it deems relevant for the user. The third information source is an *endogenous* one and consists of the possibility to adopt assumptions by default. An agent may for instance assume by default that its user will be interested in incoming email concerning a workshop on information agents (for example since it knows that the user is interested in agents in general). All these kinds of information acquisition are formalised by so-called *informative* actions, to be performed by the agents. Before we introduce these actions, we first elaborate on the degree of credibility of, and the possibility of conflicts between, information.

5.2 A classification of information

In any situation in which an agent is busy acquiring information, the possibility of conflicts exists. That is, newly acquired information may contradict already present, previously acquired information. Since the information of (rational) agents is in general demanded to be non-absurd, it has to be decided how these information conflicts are to be solved. According to the most well known paradigm to information change, the AGM axiomatisation for belief change due to Alchourrón, Gärdenfors and Makinson [1, 38, 39], priority should be given to the most recently acquired information. That is, in the case of a conflict between new, incoming information and already present information, the first should prevail. However, in the presence of various different sources of information, as is the case in our setting, the principle of priority to most recently acquired information no longer takes root. For one of the sources may be more credible than another one, a situation which applies to our framework, which should imply that information acquired from the former, more credible source should always prevail over that acquired from the latter one, regardless of the order in which the information is acquired. The strategy that we propose to replace the AGM paradigm is therefore that incoming information only causes a revision if it does not conflict with already present information of at least the same credibility; if a revision indeed takes place then it should be performed to comply as much as possible with the AGM postulates for belief revision.

To formalise these intuitive ideas we need a way to attach a degree of credibility both to the information that an agent has, and to the sources used to acquire (additional) information. The credibility ordering of the three sources of information as we propose it, is a rather straightforward one: roughly speaking, observations are more credible than communication and default assumptions, and communication is more credible than default reasoning (more details can be found in Sections 5.5 to 5.7). In order to impose a credibility ordering on the information of an agent, we propose to structure this information into four sets, situated within each other. The innermost of these sets contains the *knowledge* of the agent. Knowledge can be seen as the most credible kind of information available to the agent. This information is in particular such that it is never modified as the result of the execution of an informative action. The set directly encompassing the agent's knowledge contains the *observational* beliefs of the agent. These are the beliefs that an agent has on the ground of its knowledge and the observations it has made. The third set contains the *communicational* beliefs of an agent. These are the combined beliefs that it either knows or acquired through observations and/or communication. The outermost set contains the *default beliefs*. These are the beliefs for which application of a default may have been a necessary condition. The credibility of a belief formula is determined by the smallest belief set that it is a member of, i.e. formulae that are known have the highest credibility whereas formulae that are believed by default have

the lowest one.

In the spirit of Hintikka's representation of belief [48], we use operators $\mathbf{B}^x\_$ to refer to the various notions of belief that we formalise. The formula $\mathbf{B}^o_i\varphi$ denotes that φ belongs to the observational beliefs of agent i, $\mathbf{B}^c_i\varphi$ denotes that φ is one of i's communicational beliefs, and $\mathbf{B}^d_i\varphi$ denotes that i believes φ by default. In addition to these doxastic operators, the language L^{I} furthermore contains an operator $\mathbf{D}\_,\_$, which is used to model the dependence or authorisation relations between agents: $\mathbf{D}_{i,j}\varphi$ denotes that agent i accepts j as an authority on the subject φ. The relevance of this operator will be made clear in Section 5.6. The class Ac^{I} of actions is built up from the core constructors and three new constructors, viz. $\mathtt{observe}\,\_$, $\mathtt{inform}\,(\_,\_)$ and $\mathtt{try\_jump}\,\_$. Whenever f is a propositional formula, $\mathtt{observe}\,f$ is the action consisting of observing whether f holds, and $\mathtt{try\_jump}\,f$ is the action of adopting f by default. For $f \in \mathrm{L}_0$ and $j \in \mathrm{A}$, the action $\mathtt{inform}\,(f,j)$ formalises the act of telling agent j that f holds. It is obvious that if one is to model realistic communication, more than one agent should be present, which explains the constraint that A contains at least two elements in Definition 5.1. The demand for finiteness is explained in Remark 5.32.

5.1. Definition. To define the language L^{I}, it is demanded that the set A of agents, on which L^{I} is founded, is finite and contains at least two elements. The alphabet is extended with the doxastic operators $\mathbf{B}^o\_$, $\mathbf{B}^c\_$ and $\mathbf{B}^d\_$, the dependence operator $\mathbf{D}\_,\_$ and the action constructors $\mathtt{observe}\,\_$, $\mathtt{try\_jump}\,\_$ and $\mathtt{inform}\,(\_,\_)$. The language L^{I} is the smallest superset of Π such that the core clauses are validated and furthermore

- if $\varphi \in \mathrm{L}^{\mathrm{I}}$ and $i \in \mathrm{A}$ then $\mathbf{B}^o_i\varphi \in \mathrm{L}^{\mathrm{I}}$, $\mathbf{B}^c_i\varphi \in \mathrm{L}^{\mathrm{I}}$ and $\mathbf{B}^d_i\varphi \in \mathrm{L}^{\mathrm{I}}$
- if $\varphi \in \mathrm{L}^{\mathrm{I}}$ and $i \in \mathrm{A}$, $j \in \mathrm{A}$ then $\mathbf{D}_{i,j}\varphi \in \mathrm{L}^{\mathrm{I}}$

The class Ac^{I} is the smallest superset of At closed under the core clauses and such that

- if $f \in \mathrm{L}_0$ and $j \in \mathrm{A}$ then $\mathtt{observe}\,f \in \mathrm{Ac}^{\mathrm{I}}$, $\mathtt{inform}\,(f,j) \in \mathrm{Ac}^{\mathrm{I}}$ and $\mathtt{try\_jump}\,f \in \mathrm{Ac}^{\mathrm{I}}$

5.2. Remark. Both for reasons of practical convenience and notational uniformity, we sometimes use $\mathbf{B}^k\_$, which may be read as denoting *known beliefs*, to represent $\mathbf{K}\_$.

The restriction to propositional formulae as appearing as arguments of the action constructors modelling information acquisition is dictated by the limits of the AGM paradigm for belief revision. For execution of an informative action will in general cause the beliefs of the agent that executes the action to be revised. The AGM paradigm describes how non-consecutive changes of belief are to be implemented for propositional formulae [1, 38, 39]; several extensions and modifications of the AGM axiomatisation deal with iterated changes of belief with propositional formulae [22, 77, 81]. Changes of belief with other formulae, and in particular with epistemic or doxastic formulae, are not equally well understood. It is for instance not at all clear what it means to revise

the beliefs of some agent i with the formula $p \wedge \neg \mathbf{B}_i^o p$: should or shouldn't i believe $p \wedge \neg \mathbf{B}_i^o p$ as the result of a revision of its beliefs with this formula? For believing the formula would imply that the agent both believes that it believes p and believes that it does not believe p, which would render its beliefs inconsistent[1]. Having said so, it must be remarked that a restriction to propositional formulae is for our framework in fact too severe: formulae like $\mathbf{A}_i a$ or $\langle \mathrm{do}_i(a_1; a_2) \rangle (p \wedge q)$, or any other formula of which truth does not depend on informational attitudes, do not cause any problems. Both for reasons of convenience and since an extension to this kind of formulae is not substantially different, we have decided to maintain the restriction to propositional formulae.

The models that are used to interpret formulae from L^{I} contain, besides the core elements, functions to interpret the various doxastic operators and the dependence operator. By imposing certain constraints upon the functions interpreting the doxastic operators, we ensure that these operators validate certain intuitive desiderata. Most of these desiderata stem from our view on knowledge and belief, which is to a large extent in line with the predominant one in AI and computer science. That is, the main difference between knowledge and belief is that the latter notion is not necessarily veridical, but validates the weaker demand of consistency, or non-absurdness. Hence an agent does not believe inconsistent (absurd) formulae, and has positive and negative introspection on its beliefs. Although beliefs in general are not veridical, we feel that the observational beliefs of an agent are. For these are the things that the agent believes because it saw them with its own eyes, and, assuming that the agent's eyesight is reliable, are therefore certainly true. Hence the observational beliefs of agents are not only consistent, but even veridical[2]. The other beliefs of an agent are not veridical: an agent may tell another agent things that turn out to be false (even though the first agent believed them to be true), and beliefs adopted by default are not necessarily true, by the very nature of defaults.

To formalise all of these desiderata we propose a semantics for the doxastic operators which is based on the use of so-called *belief clusters* [94]. In essence, belief clusters are nothing but (sub)sets of possible worlds that together constitute a designated body of belief of an agent. In the special way that we use them, belief clusters are sets of worlds, situated within each other, that each represent a set of beliefs of a certain credibility. A

[1] Formulae of this kind are considered by Thijsse ([125], pp. 131-132) to represent non-contradictory sentences but contradictory utterances. This implies that although the sentence in itself is consistent, it is not consistent to believe the sentence.

[2] This veridicality of observational belief, which turns it *de facto* into a kind of knowledge, raises the philosophical question whether there is room for two kinds of knowledge, living alongside each other. In our opinion this room indeed exists. The argument supporting this view dates back to Kant [66], who postulated that knowledge can be distinguished in *a priori* knowledge, belonging to pure reason, and *a posteriori* knowledge, which is acquired through experience. Adopting this point of view, our kind of knowledge can be seen as *a priori* while observational beliefs denote a kind of *a posteriori* knowledge.

formula is believed with a certain credibility iff it holds at all the possible worlds in the associated belief cluster, i.e. the doxastic operators are interpreted as necessity operators over their associated belief clusters. Corresponding to the idea that knowledge is the most credible kind of belief, the knowledge cluster is the largest, outmost one. The other belief clusters are successively situated inside the knowledge cluster. That is, the knowledge cluster may contain several observational belief clusters, each corresponding with a certain observation that the agent has made, inside of which a communicational belief and a default belief cluster are situated. We propose that each observational belief cluster contains a unique communicational and default belief cluster. The reason for the non-uniqueness of the observational belief clusters lies in the history of origin of these clusters. For the idea is that these clusters come forth from observations that the agent has made, and since the effect of observations depends on the state of the world, observations in different states of affairs may result in different observational belief clusters. The uniqueness of the communicational and default belief clusters stems from the fact that these beliefs have no objective standard to which they are measured. Therefore, it does not seem to make much sense to let communication or default reasoning result in different belief clusters rather than one cluster. Having said so, it must be remarked that in Section 5.9 we propose one possible interpretation of multiple communicational belief clusters.

5.3. Definition. A model M for the language L^{I} is a tuple containing the core elements, three functions $\mathrm{B}^o : \mathrm{A} \to \wp(\mathrm{S} \times \mathrm{S})$, $\mathrm{B}^c : \mathrm{A} \times \mathrm{S} \to \wp(\mathrm{S})$ and $\mathrm{B}^d : \mathrm{A} \times \mathrm{S} \to \wp(\mathrm{S})$ which yield the various sets of doxastic alternatives of an agent in a state, and a function $\mathrm{D} : \mathrm{A} \times \mathrm{A} \to \mathrm{S} \to \wp(\mathrm{L}^{\mathrm{I}})$, used to interpret dependence relations. The B-functions are such that for all $i \in \mathrm{A}$ and $s, s' \in \mathrm{S}$:

- $\mathrm{B}^o(i)$ is an equivalence relation
- $\mathrm{B}^d(i, s) \neq \emptyset$
- $\mathrm{B}^d(i, s) \subseteq \mathrm{B}^c(i, s) \subseteq [s]_{\mathrm{B}^o(i)} \subseteq [s]_{\mathrm{R}(i)}$
- if $s' \in [s]_{\mathrm{B}^o(i)}$ then $\mathrm{B}^c(i, s') = \mathrm{B}^c(i, s)$ and $\mathrm{B}^d(i, s') = \mathrm{B}^d(i, s)$

where $[s]_{\mathrm{B}^o(i)}$ is defined analogously to $[s]_{\mathrm{R}(i)}$.

As hinted at above, the doxastic operators are interpreted as necessity operators over their associated belief clusters.

5.4. Definition. The binary relation $\models^{\mathrm{I}}$ between a formula from L^{I} and a pair M, s consisting of a model M for L^{I} and a state s in M is for doxastic formulae defined by:

$$\begin{aligned}
\mathrm{M}, s \models^{\mathrm{I}} \mathbf{B}^o_i\varphi &\Leftrightarrow \forall s' \in \mathrm{S}((s, s') \in \mathrm{B}^o(i) \Rightarrow \mathrm{M}, s' \models^{\mathrm{I}} \varphi) \\
\mathrm{M}, s \models^{\mathrm{I}} \mathbf{B}^c_i\varphi &\Leftrightarrow \forall s' \in \mathrm{S}(s' \in \mathrm{B}^c(i, s) \quad \Rightarrow \mathrm{M}, s' \models^{\mathrm{I}} \varphi) \\
\mathrm{M}, s \models^{\mathrm{I}} \mathbf{B}^d_i\varphi &\Leftrightarrow \forall s' \in \mathrm{S}(s' \in \mathrm{B}^d(i, s) \quad \Rightarrow \mathrm{M}, s' \models^{\mathrm{I}} \varphi)
\end{aligned}$$

When interpreting the doxastic operators as in Definition 5.4 for the models given in Definition 5.3, these operators indeed validate the desiderata formulated above. It is in particular the case that these operators may be compared according to credibility (Proposition 5.5), and that they validate the desired axiomatisations (Proposition 5.6).

5.5. Proposition. *Define the ordering* $>$ *on informational operators by* $\mathbf{B}_i^k > \mathbf{B}_i^o > \mathbf{B}_i^c > \mathbf{B}_i^d$*, and let* $\geq$ *be the reflexive, transitive closure of* $>$*. Then for all* $\mathbf{X}, \mathbf{Y} \in \{\mathbf{B}_i^k, \mathbf{B}_i^o, \mathbf{B}_i^c, \mathbf{B}_i^d\}$*, and for all* $\varphi \in \mathrm{L}^{\mathrm{I}}$ *we have that if* $\mathbf{X} \geq \mathbf{Y}$ *then* $\models^{\mathrm{I}} \mathbf{X}\varphi \to \mathbf{Y}\varphi$*.*

5.6. Proposition. *Let* $\mathbf{X} \in \{\mathbf{B}_i^o, \mathbf{B}_i^c, \mathbf{B}_i^d\}$*. For all* $\varphi \in \mathrm{L}^{\mathrm{I}}$ *we have:*

1. $\models^{\mathrm{I}} \mathbf{X}(\varphi \to \psi) \to (\mathbf{X}\varphi \to \mathbf{X}\psi)$ *K*
2. $\models^{\mathrm{I}} \neg(\mathbf{X}\varphi \wedge \mathbf{X}\neg\varphi)$ *D*
3. $\models^{\mathrm{I}} \mathbf{B}_i^o\varphi \to \varphi$ *T*
4. $\models^{\mathrm{I}} \mathbf{X}\varphi \to \mathbf{X}\mathbf{X}\varphi$ *4*
5. $\models^{\mathrm{I}} \neg\mathbf{X}\varphi \to \mathbf{X}\neg\mathbf{X}\varphi$ *5*
6. $\models^{\mathrm{I}} \varphi \Rightarrow \models^{\mathrm{I}} \mathbf{X}\varphi$ *N*

Since the doxastic operators model essentially different informational attitudes, one does obviously not want these operators to collapse. For a certain class of formulae, however, certain combinations of doxastic operators do collapse. To formalise this property we extend the notion of i-doxastic sequenced formulae as introduced by Van der Hoek [51] for a system containing knowledge and (plain) belief to deal with the doxastic operators that we consider.

5.7. Definition. A formula $\chi \in \mathrm{L}^{\mathrm{I}}$ is i-doxastic sequenced if there is some $\varphi \in \mathrm{L}^{\mathrm{I}}$ and operators $\mathbf{X}_1, \ldots, \mathbf{X}_m \in \{\mathbf{B}_i^o, \mathbf{B}_i^c, \mathbf{B}_i^d, \neg\mathbf{B}_i^o, \neg\mathbf{B}_i^c, \neg\mathbf{B}_i^d\}$ and $m > 0$ such that $\chi = \mathbf{X}_1 \ldots \mathbf{X}_m\varphi$.

5.8. Proposition. *For all* i*-doxastic sequenced formulae* $\chi \in \mathrm{L}^{\mathrm{I}}$*, for all* $\varphi \in \mathrm{L}^{\mathrm{I}}$ *and for all* $\mathbf{X} \in \{\mathbf{B}_i^o, \mathbf{B}_i^c, \mathbf{B}_i^d\}$ *we have:*

1. $\models^{\mathrm{I}} \mathbf{X}\chi \leftrightarrow \chi$
2. $\models^{\mathrm{I}} \mathbf{X}\mathbf{K}_i\varphi \leftrightarrow \mathbf{K}_i\varphi$
3. $\models^{\mathrm{I}} \mathbf{X}\neg\mathbf{K}_i\varphi \leftrightarrow \neg\mathbf{K}_i\varphi$
4. $\mathbf{X}\varphi \to \mathbf{K}_i\mathbf{X}\varphi$ *is not for all* $\varphi \in \mathrm{L}^{\mathrm{I}}$ *valid*
5. $\neg\mathbf{X}\varphi \to \mathbf{K}_i\neg\mathbf{X}\varphi$ *is not for all* $\varphi \in \mathrm{L}^{\mathrm{I}}$ *valid*

The first item of Proposition 5.8 implies that i-doxastic sequenced formulae are observationally believed by agent i iff they are communicationally believed iff they are believed by default; items 2 and 3 state that the same holds for formulae prefixed with $\mathbf{K}_i$ or $\neg\mathbf{K}_i$. The last two items of Proposition 5.8 state that the propositions formulated

in items 2 and 3 do not hold when swapping $\mathbf{X}$ and $\mathbf{K}_i$ in the left-hand side of the equivalence. That is, it is for instance not the case that whenever agent i observationally believes φ it also knows that it does so. For observational beliefs this is obvious: since $\mathbf{B}_i^o$ is veridical, it would imply the validity of $\mathbf{B}_i^o\psi \rightarrow \mathbf{K}_i\psi$, which would collapse knowledge and observational beliefs. That these non-validities then also apply for communicational beliefs and default beliefs, i.e. agents do not necessarily know that they believe something by default if they do so, is a consequence of the formal implementation of the various doxastic operators. For it is demanded that each observational belief cluster contains a unique communicational and default belief cluster, but since various observational belief clusters may be situated within one knowledge cluster, it does not follow that each knowledge cluster contains its own unique communicational and default belief cluster. The last two items of Proposition 5.8 are perhaps not for everyone completely acceptable from an intuitive point of view, i.e. it seems quite reasonable to demand agents to know of all their beliefs. However, as Corollary 5.9 states, agents are aware of all their information at the level of their observational beliefs, which is a level with a considerable degree of credibility attached to it.

5.9. Corollary. *For all* $\varphi \in \mathrm{L}^{\mathrm{I}}$ *and for all* $\mathbf{X} \in \{\mathbf{B}_i^k, \mathbf{B}_i^o, \mathbf{B}_i^c, \mathbf{B}_i^d\}$ *we have:*

- $\models^{\mathrm{I}} \mathbf{X}\varphi \leftrightarrow \mathbf{B}_i^o\mathbf{X}\varphi$
- $\models^{\mathrm{I}} \neg\mathbf{X}\varphi \leftrightarrow \mathbf{B}_i^o\neg\mathbf{X}\varphi$

Given the four modal operators $\mathbf{B}_i^k, \mathbf{B}_i^o, \mathbf{B}_i^c, \mathbf{B}_i^d$ and the credibility ordering between them, it is possible to model exactly nine different informational attitudes of a given agent with regard to a given formula. If we define for $x \in \{k, o, c, d\}$, the operators $\mathbf{Bwhether}_i^x\varphi$, representing the fact that agent i believes whether φ at level x, and $\mathbf{Ignorant}_i^x\varphi$, representing the fact that agent i is ignorant with regard to φ on the level x, by

- $\mathbf{Bwhether}_i^x\varphi \triangleq \mathbf{B}_i^x\varphi \vee \mathbf{B}_i^x\neg\varphi$
- $\mathbf{Ignorant}_i^x\varphi \triangleq \neg\mathbf{Bwhether}_i^x\varphi$

the nine possible informational attitudes with respect to a formula φ are the following:

1. $\mathbf{B}_i^k\varphi$ — 'i knows φ'
2. $\mathbf{B}_i^k\neg\varphi$ — 'i knows $\neg\varphi$'
3. $\mathbf{Ignorant}_i^k\varphi \wedge \mathbf{B}_i^o\varphi$ — 'i saw φ'
4. $\mathbf{Ignorant}_i^k\varphi \wedge \mathbf{B}_i^o\neg\varphi$ — 'i saw $\neg\varphi$'
5. $\mathbf{Ignorant}_i^o\varphi \wedge \mathbf{B}_i^c\varphi$ — 'i was told φ'
6. $\mathbf{Ignorant}_i^o\varphi \wedge \mathbf{B}_i^c\neg\varphi$ — 'i was told $\neg\varphi$'
7. $\mathbf{Ignorant}_i^c\varphi \wedge \mathbf{B}_i^d\varphi$ — 'i believes φ by default'
8. $\mathbf{Ignorant}_i^c\varphi \wedge \mathbf{B}_i^d\neg\varphi$ — 'i believes $\neg\varphi$ by default'
9. $\mathbf{Ignorant}_i^d\varphi$ — 'i is completely ignorant wrt φ'

The intuitive interpretation of the various formalised informational attitudes is given at the right-hand side. For example, the formula $\mathbf{Ignorant}_i^o\varphi \wedge \mathbf{B}_i^c\varphi$, which states that agent i does not observationally believe whether φ holds but does so on the level of its communicational beliefs, is taken to represent that i was told that φ holds, i.e. some other agent communicated to i that φ is the case. To enhance the reasoning with and about the agents' informational attitudes we introduce the following predicates by definitional abbreviation:

5.10. Definition. For $i \in \mathrm{A}$ and $\varphi \in \mathrm{L}^{\mathrm{I}}$ we define:

- $\mathbf{Saw}_i\varphi \triangleq \mathbf{Ignorant}_i^k\varphi \wedge \mathbf{B}_i^o\varphi$
- $\mathbf{Heard}_i\varphi \triangleq \mathbf{Ignorant}_i^o\varphi \wedge \mathbf{B}_i^c\varphi$
- $\mathbf{Jumped}_i\varphi \triangleq \mathbf{Ignorant}_i^c\varphi \wedge \mathbf{B}_i^d\varphi$

As mentioned above, the $\mathbf{Heard}_i$ operator does not formalise hearing *per se*, but rather *believing on the basis of being told*. For one cannot prevent an agent from being told inconsistencies, or formulae that it already knew or believed observationally. However, hearing these formulae does not modify the beliefs of the agent that are grounded in the things that it has been told.

5.11. Remark. Note that the definition of $\mathbf{Ignorant}_i^x$ is such that $\mathbf{Ignorant}_i^x\varphi$ and $\mathbf{Ignorant}_i^x\neg\varphi$ are equivalent notions, for $x \in \{d, c, o, k\}$. Note furthermore that $\mathbf{Saw}_i\varphi$, $\mathbf{Heard}_i\varphi$ and $\mathbf{Jumped}_i\varphi$ could equivalently — and less complex — be defined as $\mathbf{B}_i^o\varphi \wedge \neg\mathbf{B}_i^k\varphi$, $\mathbf{B}_i^c\varphi \wedge \neg\mathbf{B}_i^o\varphi$ and $\mathbf{B}_i^d\varphi \wedge \neg\mathbf{B}_i^c\varphi$, respectively. To emphasise that the agent is genuinely *ignorant* at higher levels, we have chosen to include the $\mathbf{Ignorant}_i^x$ operator in the definition of the derived belief operators, even though we feel free to use the condensed version whenever convenient.

To characterise the derived informational operators introduced in Definition 5.10 we investigate in a structured way which of the axioms of modal logic given by Chellas [19] are validated by the various operators.

5.12. Proposition. *Let $\varphi, \psi \in \mathrm{L}^{\mathrm{I}}$ and $i \in \mathrm{A}$ be arbitrary. Let $\mathbf{X}$ be in the set $\{\mathbf{Heard}_i, \mathbf{Jumped}_i\}$, $\mathbf{Y} \in Bel = \{\mathbf{Saw}_i, \mathbf{Heard}_i, \mathbf{Jumped}_i\}$, and let $\mathbf{Z} \in Bel \cup \{\mathbf{B}_i^k\}$. Define the ordering $\geq'$ to be the reflexive and transitive closure of $>'$ with $\mathbf{B}_i^k >' \mathbf{Saw}_i >' \mathbf{Heard}_i >' \mathbf{Jumped}_i$. Then we have:*

1. $\models^{\mathrm{I}} \mathbf{Y}\varphi \wedge \mathbf{Y}(\varphi \to \psi) \to \mathbf{Y}\psi$ *K*
2. $\models^{\mathrm{I}} \neg(\mathbf{Y}\varphi \wedge \mathbf{Y}\neg\varphi)$ *D*
3. $\models^{\mathrm{I}} \mathbf{Saw}_i\varphi \to \varphi$ *T*
4. $\models^{\mathrm{I}} \mathbf{Y}\varphi \to \mathbf{B}_i^o\mathbf{Y}\varphi$ *4*
5. $\models^{\mathrm{I}} \mathbf{X}\varphi \to \neg\mathbf{X}\mathbf{X}\varphi$ *4'*

6. $\models^{\mathrm{I}} \mathbf{Saw}_i\varphi \wedge \mathbf{M}_i\neg\mathbf{Saw}_i\varphi \leftrightarrow \mathbf{Saw}_i\mathbf{Saw}_i\varphi$ *4"*
7. $\models^{\mathrm{I}} \neg\mathbf{Y}\varphi \rightarrow \mathbf{B}^o_i\neg\mathbf{Y}\varphi$ *5*
8. $\models^{\mathrm{I}} \neg\mathbf{X}\varphi \rightarrow \neg\mathbf{X}\neg\mathbf{X}\varphi$ *5'*
9. $\models^{\mathrm{I}} \neg\mathbf{Saw}_i\varphi \wedge \mathbf{M}_i\mathbf{Saw}_i\varphi \leftrightarrow \mathbf{Saw}_i\neg\mathbf{Saw}_i\varphi$ *5"*
10. $\models^{\mathrm{I}} (\mathbf{Y}\varphi \wedge \mathbf{Y}\psi) \rightarrow \mathbf{Y}(\varphi \wedge \psi)$ *C*
11. $\mathbf{Y}(\varphi \wedge \psi) \rightarrow (\mathbf{Y}\varphi \wedge \mathbf{Y}\psi)$ *is not for all* $\varphi, \psi \in \mathrm{L}^{\mathrm{I}}$ *valid* *M*
12. $\models^{\mathrm{I}} \mathbf{Y}(\varphi \wedge \psi) \rightarrow (\bigvee_{\mathbf{Z} \geq' \mathbf{Y}} \mathbf{Z}\varphi \wedge \bigvee_{\mathbf{Z} \geq' \mathbf{Y}} \mathbf{Z}\psi) \wedge (\mathbf{Y}\varphi \vee \mathbf{Y}\psi)$ *M'*
13. $\models^{\mathrm{I}} \varphi \Rightarrow \models^{\mathrm{I}} \neg\mathbf{Y}\varphi$ *N*
14. *not for all* $\varphi, \psi \in \mathrm{L}^{\mathrm{I}}$ *does* $\models^{\mathrm{I}} \varphi \rightarrow \psi$ *imply* $\models^{\mathrm{I}} \mathbf{Y}\varphi \rightarrow \mathbf{Y}\psi$ *RM*
15. $\models^{\mathrm{I}} \varphi \rightarrow \psi \Rightarrow \models^{\mathrm{I}} \mathbf{Y}\varphi \rightarrow \bigvee_{\mathbf{Z} \geq' \mathbf{Y}} \mathbf{Z}\psi$ *RM'*
16. $\models^{\mathrm{I}} \varphi \leftrightarrow \psi \Rightarrow \models^{\mathrm{I}} \mathbf{Y}\varphi \leftrightarrow \mathbf{Y}\psi$ *RE*

That the K-axiom holds for all three derived belief operators might look somewhat surprising at first sight. However, validity of this axiom has everything to do with the *rationality* of agents. Take the example of a rational agent grounding its belief in both φ and $\varphi \rightarrow \psi$ in its observations. Being the rational creature that it is, it is obvious that it in any case observationally believes ψ. Again given the rationality of the agent it cannot attach a higher credibility to ψ than to $\varphi \rightarrow \psi$: for the latter is implied by the former, and therefore the credibility attached to it is at least the credibility attached to ψ. Hence both ψ and $\varphi \rightarrow \psi$ are believed with the same strength, which explains the validity of the K-axiom. Beliefs are consistent (the D-axiom), and beliefs that are grounded in observations are furthermore veridical (the T-axiom). The first property is highly desirable for rational agents, the latter property is the essential characteristic of beliefs acquired through observations (cf. [7], p. 12). That in general agents do not have positive (4,4',4") and negative (5,5'5") introspection on the derived belief operators is a direct consequence of Corollary 5.9. The validities given in the items 6 and 9 are fairly conspicuous. They state that in certain circumstances agents may have positive and negative introspection on the beliefs they acquired through observation, i.e. it is possible that an agent saw that it saw φ because it saw φ. Although these validities are possibly not completely acceptable from an intuitive point of view, they are easily explained from a technical point of view, viz. the definition of the derived operators and the properties of the primitive doxastic operators. For whenever an agent i saw φ to hold, then by item 4 it observationally believes that it did so, i.e. $\mathbf{B}^o_i\mathbf{Saw}_i\varphi$ holds. If the agent considers it furthermore epistemically possible that it did not see φ, then both $\mathbf{B}^o_i\mathbf{Saw}_i\varphi$ and $\neg\mathbf{K}_i\mathbf{Saw}_i\varphi$ hold which suffices to conclude $\mathbf{Saw}_i\mathbf{Saw}_i\varphi$. Analogous arguments can be given to explain the reverse implication of item 6 and the equivalence of item 9. We suggest that when appearing in formulae as given in items 6 and 9, the $\mathbf{Saw}_i$ operator should not be interpreted as 'having seen' but instead simply as 'observationally believing but not knowing'. All derived belief operators validate the C-axiom, none validates the

M-axiom, but all validate a variant of the M-axiom. The validity given in item 12 states that if an agent believes a conjunction with a certain credibility and not with a higher one, then it believes both conjuncts at least with the same credibility as attached to the conjunction, while at least one of the conjuncts is believed with exactly the same credibility. For example, if an agent believes $\varphi \wedge \psi$ since it is told that this holds, then it knows, saw, or was told both φ and ψ, while it believes either φ or ψ on the ground of it being told this to be true. None of the operators satisfy necessitation (the N-rule), the reason for this being the fact that valid propositions are already known, and are therefore never grounded in observations, communication or default jumps. The derived belief operators are in general not monotonic (the RM-rule), but satisfy some kind of 'upward monotonicity' (the RM'-rule), corresponding to the idea that whenever some proposition is believed to a certain degree, weaker propositions might already be believed to a higher degree (the degenerate case being one where the weaker propositions are validities that are necessarily known to the agent). Finally, all derived belief operators are closed under equivalence of (believed) propositions. Although it is sometimes argued against this property (cf. [7], p. 18), it seems harmless for artificial agents with perfect reasoning capacities.

5.3 Informative actions as model-transformers

To treat informative actions as genuine, fully-fledged actions, one has to decide both upon an informal description of the result, opportunity and ability of these actions, i.e. what is it that constitutes these notions, and upon a formal interpretation in terms of our framework. The one notion that is most easily described is probably that of the result. For although the results of different informative actions are different in that execution of an `observe` f action will in general affect all of the beliefs of an agent whereas executing `try_jump` f affects only the agent's default beliefs, their common characteristic is that they cause a modification, i.e. revision, of the agent's beliefs. With regard to opportunity and ability these actions are much less similar. Therefore we focus at first on the result of informative actions, leaving the informal description as well as the formal definition of ability and opportunity to the sections dealing with the respective actions.

If the result of execution of an informative action is somehow to modify the beliefs of an agent, we have to ensure that this is formalised in our semantics. Our guideline throughout defining the semantics is that whenever execution of an informative action should cause a change in the agent's information, or better a revision of the agent's beliefs, then it should do so in compliance with the AGM postulates for belief revision. More in particular, we consider the notion of *minimal change* as sacrosanct, i.e. the performance of an informative action should cause only those changes that are necessary for the action to be informative. We can imagine at least two ways of defining an adequate

formal semantics that complies with this guideline. The first would be to interpret informative actions as transitions to states that minimally differ from the starting state, while the agent's beliefs have been revised appropriately. Given the fact that agents may in principle perform arbitrary informative actions, this would probably require some kind of fullness-condition to be imposed on the models. That is, for every agent i and every state s in a model, every possible doxastic state, i.e. every state of the agent's beliefs that could possibly result from some change in the beliefs of i in s in compliance with the AGM postulates, should be present somewhere, i.e at some state s', in the model. Although it could be possible to formally define this fullness-condition, it would cause the models to be enormous, with a gigantic overhead of states that are possibly never used in any reasoning of or about the agents. And even though the computational intractability that would possibly be involved with this kind of enormous models is not our primary concern, the ponderousness of this solution is. The second way of defining adequate semantics for informative actions — which is the one that we will pursue — is based on the idea that every state in a model is most similar to itself. Hence, if one wants informative actions to affect the belief sets of an agent in some state while causing as little change as possible, it might be a good idea to change just the doxastic state of the agent, embodied through the B-functions, while leaving the state in itself unaffected. In this way one can easily ensure that informative actions result in the appropriate changes in the agent's beliefs, while at the same time changing as little as possible. Note that this suggested interpretation of informative actions as *transforming models* is a generalisation of the standard paradigm of Propositional Dynamic Logic, in which actions are interpreted as causing *transitions between states*.

The interpretation of informative actions as we define it later on is such that the models resulting from execution of an informative action in some state of some model M resemble M in all aspects except (possibly) the B-functions. To account for this property formally, we introduce the class of models similar to a given model.

5.13. Definition. Let $\mathrm{M} \in \mathbf{M}^{\mathrm{I}}$ be some model for L^{I}. The class $\mathrm{M}^{\mathrm{I}}_{\sim} \subseteq \mathbf{M}^{\mathrm{I}}$ contains all models that (possibly) differ from M only in the B-functions.

To allow ordinary, mundane actions to occur alongside informative actions, we combine the state-transition and the model-transforming interpretation in our definition of the functions r^{I} and c^{I} used to interpret the non-core formulae. In Definition 5.3 we present that part of r^{I} and c^{I} that deals with the core actions; the more interesting part of the definition of r^{I} and c^{I}, i.e. the part dealing with informative actions, is presented in Sections 5.5 through 5.7.

5.14. Definition. The binary relation $\models^{\mathrm{I}}$ between a formula $\varphi \in \mathrm{L}^{\mathrm{I}}$ and a pair M, s consisting of a model M for L^{I} and a state s in M is for dynamic and ability formulae defined by:

$$\begin{array}{ll} \mathrm{M}, s \models^{\mathrm{I}} \langle \mathrm{do}_i(\alpha)\rangle\varphi & \Leftrightarrow \exists \mathrm{M}', s' \in \mathrm{M}^{\mathrm{I}}_{\sim} \times \mathrm{S}(\mathrm{M}', s' = \mathtt{r}^{\mathrm{I}}(i,\alpha)(\mathrm{M}, s) \,\&\, \mathrm{M}', s' \models^{\mathrm{I}} \varphi) \\ \mathrm{M}, s \models^{\mathrm{I}} \mathbf{A}_i\alpha & \Leftrightarrow \mathtt{c}^{\mathrm{I}}(i,\alpha)(\mathrm{M}, s) = \mathbf{1} \end{array}$$

where $\mathtt{r}^{\mathrm{I}}$ and $\mathtt{c}^{\mathrm{I}}$ are for the core actions defined by:

$$\begin{array}{ll}
\mathtt{r}^{\mathrm{I}} & : \ \mathrm{A} \times \mathrm{Ac}^{\mathrm{I}} \to (\mathrm{M}^{\mathrm{I}}_{\sim} \times \mathrm{S})^{\cdot} \to (\mathrm{M}^{\mathrm{I}}_{\sim} \times \mathrm{S})^{\cdot} \\
\mathtt{r}^{\mathrm{I}}(i,a)(\mathrm{M}', s) & = \mathrm{M}', \mathtt{r}_0(i,a)(s) \text{ if } \mathtt{r}_0(i,a)(s) \neq \emptyset \\
& = \emptyset \text{ otherwise} \\
\mathtt{r}^{\mathrm{I}}(i, \mathtt{confirm}\,\varphi)(\mathrm{M}', s) & = \mathrm{M}', s \text{ if } \mathrm{M}', s \models^{\mathrm{I}} \varphi \\
& = \emptyset \text{ otherwise} \\
\mathtt{r}^{\mathrm{I}}(i, \alpha_1; \alpha_2)(\mathrm{M}', s) & = \mathtt{r}^{\mathrm{I}}(i,\alpha_2)(\mathtt{r}^{\mathrm{I}}(i,\alpha_1)(\mathrm{M}', s)) \\
\mathtt{r}^{\mathrm{I}}(i, \mathtt{if}\,\varphi\,\mathtt{then}\,\alpha_1\,\mathtt{else}\,\alpha_2\,\mathtt{fi})(\mathrm{M}', s) & = \mathtt{r}^{\mathrm{I}}(i,\alpha_1)(\mathrm{M}', s) \text{ if } \mathrm{M}', s \models^{\mathrm{I}} \varphi \\
& = \mathtt{r}^{\mathrm{I}}(i,\alpha_2)(\mathrm{M}', s) \text{ otherwise} \\
\mathtt{r}^{\mathrm{I}}(i, \mathtt{while}\,\varphi\,\mathtt{do}\,\alpha\,\mathtt{od})(\mathrm{M}', s) & = \mathrm{M}'', s' \text{ if, for some } k \in \mathbb{N}, \\
& \quad \mathrm{M}'', s' = \mathtt{r}^{\mathrm{I}}(i, (\mathtt{confirm}\,\varphi; \alpha)^k; \mathtt{confirm}\,\neg\varphi)(\mathrm{M}', s) \\
& = \emptyset \text{ otherwise} \\
\mathtt{r}^{\mathrm{I}}(i,\alpha)(\emptyset) & = \emptyset
\end{array}$$

$$\begin{array}{ll}
\mathtt{c}^{\mathrm{I}} & : \ \mathrm{A} \times \mathrm{Ac}^{\mathrm{I}} \to (\mathrm{M}^{\mathrm{I}}_{\sim} \times \mathrm{S})^{\cdot} \to \mathsf{bool} \\
\mathtt{c}^{\mathrm{I}}(i,a)(\mathrm{M}', s) & = \mathbf{1} \text{ iff } \mathtt{c}_0(i,a)(s) = \mathbf{1} \\
\mathtt{c}^{\mathrm{I}}(i, \mathtt{confirm}\,\varphi)(\mathrm{M}', s) & = \mathbf{1} \text{ iff } \mathrm{M}', s \models^{\mathrm{I}} \varphi \\
\mathtt{c}^{\mathrm{I}}(i, \alpha_1; \alpha_2)(\mathrm{M}', s) & = \mathbf{1} \text{ iff } \mathtt{c}^{\mathrm{I}}(i,\alpha_1)(\mathrm{M}', s) = \mathbf{1} \,\& \\
& \qquad \mathtt{c}^{\mathrm{I}}(i,\alpha_2)(\mathtt{r}^{\mathrm{I}}(i,\alpha_1)(\mathrm{M}, s)) = \mathbf{1} \\
\mathtt{c}^{\mathrm{I}}(i, \mathtt{if}\,\varphi\,\mathtt{then}\,\alpha_1\,\mathtt{else}\,\alpha_2\,\mathtt{fi})(\mathrm{M}', s) & = \mathbf{1} \text{ iff } \mathtt{c}^{\mathrm{I}}(i, \mathtt{confirm}\,\varphi; \alpha_1)(\mathrm{M}', s) = \mathbf{1} \text{ or} \\
& \qquad \mathtt{c}^{\mathrm{I}}(i, \mathtt{confirm}\,\neg\varphi; \alpha_2)(\mathrm{M}', s) = \mathbf{1} \\
\mathtt{c}^{\mathrm{I}}(i, \mathtt{while}\,\varphi\,\mathtt{do}\,\alpha\,\mathtt{od})(\mathrm{M}', s) & = \mathbf{1} \text{ iff } \mathtt{c}^{\mathrm{I}}(i, (\mathtt{confirm}\,\varphi; \alpha)^k; \mathtt{confirm}\,\neg\varphi)(\mathrm{M}', s) = \mathbf{1} \\
& \qquad \text{for some } k \in \mathbb{N} \\
\mathtt{c}^{\mathrm{I}}(i,\alpha)(\emptyset) & = \mathbf{1}
\end{array}$$

For an action α from Ac^{I} it holds, just like for those from Ac in Chapter 3, that the operator $[\mathrm{do}_i(\alpha)]$, with $i \in \mathrm{A}$, is normal, i.e. this operator validates both the K-axiom and the N-rule. Thus interpreting actions as model-transformers rather than state-transitions does not affect the normality of the dynamic operators $[\mathrm{do}_i(\alpha)]$. The reason for this is in fact fairly obvious: since validity is defined as truth in *all* models of a given class, it suffices to show that execution of an action transforms a model into another (well-formed) model from that class.

5.15. Proposition. *Let $\alpha \in \mathrm{Ac}^{\mathrm{I}}$ be arbitrary. If for all $\mathrm{M} \in \mathbf{M}^{\mathrm{I}}$ with state s and all $i \in \mathrm{A}$, $\mathtt{r}^{\mathrm{I}}(i,\alpha)(\mathrm{M}, s) \in (\mathbf{M}^{\mathrm{I}} \times \mathrm{S})^{\cdot}$, then $[\mathrm{do}_i(\alpha)]$ is a normal modal operator.*

5.16. Corollary. *Let* $\alpha \in \mathrm{Ac}^{\mathrm{I}}$ *be arbitrary. If for all* $\mathrm{M} \in \mathbf{M}^{\mathrm{I}}$ *with state* s *and all* $i \in \mathrm{A}$, $\mathtt{r}^{\mathrm{I}}(i,\alpha)(\mathrm{M},s) \in (\mathbf{M}^{\mathrm{I}} \times \mathrm{S})^{\cdot}$, *then we have for all* $i \in \mathrm{A}, \varphi \in \mathrm{L}^{\mathrm{I}}$*:*

- $\models^{\mathrm{I}} [\mathrm{do}_i(\alpha)](\varphi \to \psi) \to ([\mathrm{do}_i(\alpha)]\varphi \to [\mathrm{do}_i(\alpha)]\psi)$
- $\models^{\mathrm{I}} \varphi \Rightarrow \models^{\mathrm{I}} [\mathrm{do}_i(\alpha)]\varphi$

For informative actions several interesting properties can be distinguished in addition to those given in Chapter 2. Of these the most important one is *informativeness*, which formally characterises the essence of informative actions, viz. their resulting in the acquisition of information. The property of *truthfulness* is mainly important for reliable informative actions, like observations. The idea is that the truth or falsity of the proposition that is observed is not affected by the actual act of observing. Since we more or less adopt a classical view on the world, in that all propositions are either true or false, we do not have to worry about observations at a quantum level that themselves determine the truth value of the proposition whose truth is observed.

In Proposition 5.18 we relate the additional properties introduced above with each other and those of Chapter 2.

5.17. Definition. For $\alpha \in \mathrm{Ac}^{\mathrm{I}}$ and $\varphi \in \mathrm{L}^{\mathrm{I}}$ we define:

- α is x-informative with regard to φ iff $\mathbf{F}^{\mathrm{I}} \models^{\mathrm{I}} [\mathrm{do}_i(\alpha)]\mathbf{Bwhether}_i^x \varphi$
- α is genuinely x-informative with regard to φ iff $\mathbf{F}^{\mathrm{I}} \models^{\mathrm{I}} \langle \mathrm{do}_i(\alpha)\rangle \mathbf{Bwhether}_i^x \varphi$
- α is truthful with regard to φ iff $\mathbf{F}^{\mathrm{I}} \models^{\mathrm{I}} (\varphi \to [\mathrm{do}_i(\alpha)]\varphi) \wedge (\neg\varphi \to [\mathrm{do}_i(\alpha)]\neg\varphi)$

where the right-hand side of these definitions is to be understood as a schema in i.

5.18. Proposition. *For all* $\alpha \in \mathrm{Ac}^{\mathrm{I}}$ *and* $\varphi \in \mathrm{L}^{\mathrm{I}}$ *we have:*

- *if* α *is realisable and* x*-informative with regard to* φ *then* α *is genuinely* x*-informative with regard to* φ*.*
- *if* α *is deterministic and genuinely* x*-informative with regard to* φ *then* α *is* x*-informative with regard to* φ*.*

5.4 Generalised belief revision

To formalise the actual transformation of a model due to execution of an informative action, we introduce three functions, viz. $\mathtt{revise}^x$ with $x \in \{d, c, o\}$, which are used in the definition of $\mathtt{r}^{\mathrm{I}}$ for informative actions. Informally, these functions take care of a belief revision complying with the AGM postulates, but *stretched over different levels* to ensure that the credibility ranking on the agent's beliefs is preserved. For example, consider the case where an agent observes f to be true, and hence a revision of its observational beliefs with f should take place. If observational beliefs are still to imply communicational and default beliefs, one should also revise the latter kinds of belief.

Hence the revision should not only occur at the level of the agent's observational beliefs, but should be stretched over the communicational and default beliefs as well. Essentially, a revision can be seen as consisting of dropping some information and including other information, while maintaining consistency. In our framework the information of an agent is formalised through its set of (epistemic and) doxastic alternatives. A natural implementation of the dropping and inclusion of information in terms of our framework is therefore given by the inclusion and dropping of doxastic alternatives, where the dropping of information corresponds to the inclusion of new worlds into the set of doxastic alternatives of the agent and the inclusion of (new) information corresponds to dropping some states from this set. The general idea behind the revisions formalised through the $\mathtt{revise}^x$ functions is that if incoming information f is sufficiently credible to cause a modification of some set of beliefs while not contradicting these beliefs, then it is straightforwardly added to this set by restricting the appropriate set of doxastic alternatives to those that satisfy f. In the case that f does contradict the original set of beliefs, which is, due to the veridicality of observational beliefs and the truthfulness of observations, only possible for communicational and default beliefs, then the set of beliefs is reset. That is, the new set of beliefs consists of the set of beliefs at the next level of credibility combined with f. In terms of models this corresponds to resetting the set of communicational belief alternatives to the observational belief alternatives satisfying f, and the default belief alternatives to the set of communicational belief alternatives that satisfy f. As a consequence, the function $\mathtt{revise}^x$ is for $x \in \{d, c\}$ not always defined. More in particular, it is not possible to revise the beliefs of an agent i with f if $\neg f$ is believed by i with a higher level of credibility. For this would cause the latter set of beliefs to be inconsistent, a possibility which is ruled out by Definition 5.3. When using the functions $\mathtt{revise}^x$ to interpret informative actions, i.e. in the definition of r^{I}, it is ensured that this kind of inappropriate use does not occur. That is, the definition of r^{I} is such that $\mathtt{revise}^x(i, f)$ is not applied in situations where i believes the negation of f with a credibility higher than x. Thus, even though the $\mathtt{revise}^x$ functions are not always defined, they are when used as in the definition of r^{I} for informative actions (which is given in the following sections).

Below the $\mathtt{revise}^x$ functions, for $x \in \{d, c, o\}$, are defined. In defining these functions we use $[s]^f_{\mathrm{B}^o(i)}$ to denote those states from $[s]_{\mathrm{B}^o(i)}$ that satisfy f, $[s]^{\neg f}_{\mathrm{B}^o(i)}$ to denote the set of states from $[s]_{\mathrm{B}^o(i)}$ that do not satisfy f, and $\mathrm{Cl}_{\mathrm{eq}}$ as the function that yields the Cartesian product of a given set with itself.

5.19. Definition. For $\mathrm{M} = \langle \mathrm{S}, \pi, \mathrm{R}, \mathrm{B}^o, \mathrm{B}^c, \mathrm{B}^d, \mathrm{D}, \mathtt{r}_0, \mathtt{c}_0 \rangle \in \mathbf{M}^{\mathrm{I}}$, $s \in \mathrm{S}$, $i \in \mathrm{A}$ and $f \in \mathrm{L}_0$ we define the partial functions $\mathtt{revise}^d$ and $\mathtt{revise}^c$, and the function $\mathtt{revise}^o$ as follows:

$\mathtt{revise}^d(i, f)(\mathrm{M}, s)$ is undefined if $\mathrm{M}, s \models^{\mathrm{I}} \mathbf{B}^c_i \neg f$
$\mathtt{revise}^d(i, f)(\mathrm{M}, s) = \langle \mathrm{S}, \pi, \mathrm{R}, \mathrm{B}^o, \mathrm{B}^c, \mathrm{B}^{d'}, \mathrm{D}, \mathtt{r}_0, \mathtt{c}_0 \rangle$ if $\mathrm{M}, s \not\models^{\mathrm{I}} \mathbf{B}^c_i \neg f$, where

$\mathrm{B}^{d'}(j,t) = \mathrm{B}^d(j,t)$ if $j \neq i$ or $t \notin [s]_{\mathrm{B}^o(i)}$

$$\mathrm{B}^{d'}(i,t) = \begin{cases} \mathrm{B}^d(i,t) \cap [\![f]\!] \text{ if } \mathrm{B}^d(i,t) \cap [\![f]\!] \neq \emptyset, t \in [s]_{\mathrm{B}^o(i)} \\ \mathrm{B}^c(i,t) \cap [\![f]\!] \text{ if } \mathrm{B}^d(i,t) \cap [\![f]\!] = \emptyset, t \in [s]_{\mathrm{B}^o(i)} \end{cases}$$

$\mathtt{revise}^c(i,f)(\mathrm{M},s)$ is undefined if $\mathrm{M}, s \models^{\mathrm{I}} \mathbf{B}^o_i \neg f$

$\mathtt{revise}^c(i,f)(\mathrm{M},s) = \langle \mathrm{S}, \pi, \mathrm{R}, \mathrm{B}^o, \mathrm{B}^{c'}, \mathrm{B}^{d'}, \mathrm{D}, \mathtt{r}_0, \mathtt{c}_0 \rangle$ if $\mathrm{M}, s \not\models^{\mathrm{I}} \mathbf{B}^o_i \neg f$, where

$\mathrm{B}^{c'}(j,t) = \mathrm{B}^c(j,t)$ if $j \neq i$ or $t \notin [s]_{\mathrm{B}^o(i)}$

$$\mathrm{B}^{c'}(i,t) = \begin{cases} \mathrm{B}^c(i,t) \cap [\![f]\!] \text{ if } \mathrm{B}^c(i,t) \cap [\![f]\!] \neq \emptyset, t \in [s]_{\mathrm{B}^o(i)} \\ [t]_{\mathrm{B}^o(i)} \cap [\![f]\!] \text{ if } \mathrm{B}^c(i,t) \cap [\![f]\!] = \emptyset, t \in [s]_{\mathrm{B}^o(i)} \end{cases}$$

$$\mathrm{B}^{d'}(i,t) = \begin{cases} \mathrm{B}^d(i,t) \cap [\![f]\!] \text{ if } \mathrm{B}^d(i,t) \cap [\![f]\!] \neq \emptyset, t \in [s]_{\mathrm{B}^o(i)} \\ \mathrm{B}^{c'}(i,t) \text{ if } \mathrm{B}^d(i,t) \cap [\![f]\!] = \emptyset, t \in [s]_{\mathrm{B}^o(i)} \end{cases}$$

$\mathtt{revise}^o(i,f)(\mathrm{M},s) = \langle \mathrm{S}, \pi, \mathrm{R}, \mathrm{B}^{o'}, \mathrm{B}^{c'}, \mathrm{B}^{d'}, \mathrm{D}, \mathtt{r}_0, \mathtt{c}_0 \rangle$ where

$\mathrm{B}^{o'}(j) = \mathrm{B}^o(j)$ if $j \neq i$

$\mathrm{B}^{o'}(i) = (\mathrm{B}^o(i) \setminus \mathrm{Cl}_{\mathrm{eq}}([s]_{\mathrm{B}^o(i)})) \cup \mathrm{Cl}_{\mathrm{eq}}([s]^f_{\mathrm{B}^o(i)}) \cup \mathrm{Cl}_{\mathrm{eq}}([s]^{\neg f}_{\mathrm{B}^o(i)})$

$\mathrm{B}^{c'}(j,t) = \mathrm{B}^c(j,t)$ if $j \neq i$ or $t \notin [s]_{\mathrm{B}^o(i)}$

$$\mathrm{B}^{c'}(i,t) = \begin{cases} \mathrm{B}^c(i,t) \cap [t]_{\mathrm{B}^{o'}(i)} \text{ if } \mathrm{B}^c(i,t) \cap [t]_{\mathrm{B}^{o'}(i)} \neq \emptyset, t \in [s]_{\mathrm{B}^o(i)} \\ [t]_{\mathrm{B}^{o'}(i)} \text{ if } \mathrm{B}^c(i,t) \cap [t]_{\mathrm{B}^{o'}(i)} = \emptyset, t \in [s]_{\mathrm{B}^o(i)} \end{cases}$$

$\mathrm{B}^{d'}(j,t) = \mathrm{B}^d(j,t)$ if $j \neq i$ or $t \notin [s]_{\mathrm{B}^o(i)}$

$$\mathrm{B}^{d'}(i,t) = \begin{cases} \mathrm{B}^d(i,t) \cap [t]_{\mathrm{B}^{o'}(i)} \text{ if } \mathrm{B}^d(i,t) \cap [t]_{\mathrm{B}^{o'}(i)} \neq \emptyset, t \in [s]_{\mathrm{B}^o(i)} \\ \mathrm{B}^{c'}(i,t) \text{ if } \mathrm{B}^d(i,t) \cap [t]_{\mathrm{B}^{o'}(i)} = \emptyset, t \in [s]_{\mathrm{B}^o(i)} \end{cases}$$

5.20. Proposition. *For all* $\mathrm{M} \in \mathbf{M}^{\mathrm{I}}$, $s \in \mathrm{M}$, $i \in \mathrm{A}$, $f \in \mathrm{L}_0$ *it holds for* $x \in \{c, d\}$*:*

- $\mathtt{revise}^o(i,f)(\mathrm{M},s) \in \mathbf{M}^{\mathrm{I}}$
- *if* $\mathtt{revise}^x(i,f)(\mathrm{M},s)$ *is defined, then* $\mathtt{revise}^x(i,f)(\mathrm{M},s) \in \mathbf{M}^{\mathrm{I}}$

The following example is meant to shed some more light on the, rather elaborate, definition of the $\mathtt{revise}^x$ functions.

5.21. Example. Consider the model $\mathrm{M} = \langle \mathrm{S}, \pi, \mathrm{R}, \mathrm{B}^o, \mathrm{B}^c, \mathrm{B}^d, \mathrm{D}, \mathtt{r}_0, \mathtt{c}_0 \rangle$ such that

- $\mathrm{S} = \{s, s_1, s_2, s_3, t, t_1, t_2, t_3\}$
- $\pi(p,u) = \mathbf{1}$ if $u \in \{s, s_1, s_2, s_3\}$, $\pi(p,u) = \mathbf{0}$ if $u \in \{t, t_1, t_2, t_3\}$,
  $\pi(q,u) = \mathbf{1}$ if $u \in \{s, s_1, t, t_1\}$, $\pi(q,u) = \mathbf{0}$ if $u \in \{s_2, s_3, t_2, t_3\}$,
  $\pi(r,u) = \mathbf{1}$ if $u \in \{s, s_2, t, t_2\}$, $\pi(r,u) = \mathbf{0}$ if $u \in \{s_1, s_3, t_1, t_3\}$
- $\mathrm{R}(i) = \mathrm{S} \times \mathrm{S}$
- $\mathrm{B}^o(i) = \mathrm{S} \times \mathrm{S}$
- $\mathrm{B}^c(i,u) = \{s, s_1, s_2, t, t_1, t_2\}$ for $u \in \mathrm{S}$
- $\mathrm{B}^d(i,u) = \{s\}$ for $u \in \mathrm{S}$

- $\mathtt{r}_0$ and $\mathtt{c}_0$ are arbitrary

In Figure 5.1 we graphically depict how the model M changes as the result of applying $\mathtt{revise}^o(i,p)$ to M, s. The model M', which is such that $\mathrm{M}' = \mathtt{revise}^o(i,p)(\mathrm{M}, s)$,

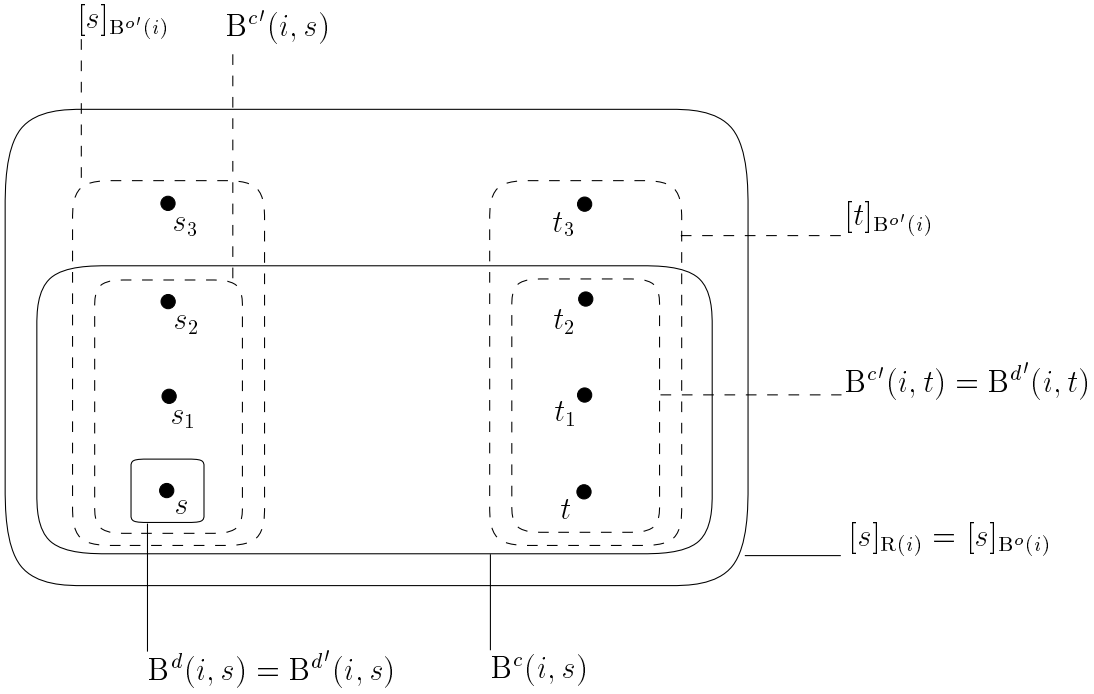

FIGURE 5.1. Revising beliefs

contains two $\mathrm{B}^{o'}(i)$-equivalence classes: one comprising all the worlds from $[s]_{\mathrm{B}^{o}(i)}$ that satisfy p, viz. $\{s, s_1, s_2, s_3\}$, and one comprising all the worlds from $[s]_{\mathrm{B}^{o}(i)}$ that do not, viz. $\{t, t_1, t_2, t_3\}$. To guarantee well-definedness of M', the set of communicational belief alternatives as it occurs in M is also split. For all the worlds u in $[s]_{\mathrm{B}^{o'}(i)}$ it holds that $\mathrm{B}^{c'}(i,u) = \{s, s_1, s_2\}$, whereas for the worlds u in $[t]_{\mathrm{B}^{o'}(i)}$ this set is given by $\{t, t_1, t_2\}$. For states in $[s]_{\mathrm{B}^{o'}(i)}$ the set of default belief alternatives for agent i is not changed as compared to M, i.e. this set is still the singleton $\{s\}$. For states $u \in [t]_{\mathrm{B}^{o'}(i)}$ a reset of default belief alternatives has taken place, which results in $\mathrm{B}^{d'}(i,u) = \mathrm{B}^{c'}(i,u) = \{t, t_1, t_2\}$.

The revision implemented by the $\mathtt{revise}^x$ functions for $x = c, d$ is what we call the *All-is-Good* (AiG) revision [84]. This kind of revision corresponds to a revision based on full meet contraction in terms of the AGM framework (cf. [1, 38]). Full meet revision constitutes the most rigorous way of revising beliefs under the AGM postulates. For instance, a revision with $\neg f$ of the beliefs of an agent that believes f results in it believing only those formulae that are somehow implied by $\neg f$. Although this rigour of

revision is in general considered unacceptable, in the following sections we will argue that for our purposes AiG revision is acceptable. The revision implemented by the $\texttt{revise}^o$ function is in fact a trivial one, viz. an expansion. This corresponds to the idea that, since observations are truthful and observational beliefs veridical, the need for a genuine revision of observational beliefs never arises. In Propositions 5.23 and 5.24 we formulate the relations between the $\texttt{revise}^x$ functions and the AGM framework. To this end we introduce some additional terminology.

5.22. Definition. For $\mathrm{M} \in \mathbf{M}^{\mathrm{I}}$, s in M, $i \in \mathrm{A}$, $f, g \in \mathrm{L}_0$ and $x \in \{o, c, d\}$ we define:

- $\mathrm{B}_x(i, \mathrm{M}, s) \triangleq \{f \in \mathrm{L}_0 \mid \mathrm{M}, s \models^{\mathrm{I}} \mathbf{B}_i^x f\}$
- $\mathrm{B}^{\perp} \triangleq \mathrm{L}_0$
- $\mathrm{B}_x^{+f}(i, \mathrm{M}, s) \triangleq \mathrm{Cn}(\mathrm{B}_x(i, \mathrm{M}, s) \cup \{f\})$
- $\mathrm{B}_x^{\star f}(i, \mathrm{M}, s) \triangleq \mathrm{B}_x(i, \mathrm{M}', s)$ if $\texttt{revise}^x(i, f)(\mathrm{M}, s) = \mathrm{M}' \in \mathbf{M}^{\mathrm{I}}$
  $\triangleq \mathrm{B}^{\perp}$ if $\texttt{revise}^x(i, f)(\mathrm{M}, s)$ is undefined
- $\mathrm{B}_x^{\star f+g}(i, \mathrm{M}, s) \triangleq \mathrm{Cn}(\mathrm{B}_x^{\star f}(i, \mathrm{M}, s) \cup \{g\})$
- $suc(d) \triangleq c$ and $suc(c) \triangleq o$

5.23. Proposition. *For all* $\mathrm{M} \in \mathbf{M}^{\mathrm{I}}$, $s \in \mathrm{M}$, $i \in \mathrm{A}$, $f, g \in \mathrm{L}_0$ *and* $x \in \{c, d\}$ *we have:*

1. $f \in \mathrm{B}_x^{\star f}(i, \mathrm{M}, s)$
2. $\mathrm{B}_x^{\star f}(i, \mathrm{M}, s) \subseteq \mathrm{B}_x^{+f}(i, \mathrm{M}, s)$
3. *If* $\neg f \notin \mathrm{B}_x(i, \mathrm{M}, s)$ *then* $\mathrm{B}_x^{+f}(i, \mathrm{M}, s) \subseteq \mathrm{B}_x^{\star f}(i, \mathrm{M}, s)$
4. $\mathrm{B}_x^{\star f}(i, \mathrm{M}, s) = \mathrm{B}^{\perp}$ *if and only if* $\mathrm{M}, s \models^{\mathrm{I}} \mathbf{B}_i^{suc(x)} \neg f$
5. *If* $\mathrm{M}, s \models^{\mathrm{I}} \mathbf{B}_i^{suc(x)}(f \leftrightarrow g)$ *then* $\mathrm{B}_x^{\star f}(i, \mathrm{M}, s) = \mathrm{B}_x^{\star g}(i, \mathrm{M}, s)$
6. $\mathrm{B}_x^{\star f \wedge g}(i, \mathrm{M}, s) \subseteq \mathrm{B}_x^{\star f+g}(i, \mathrm{M}, s)$
7. *If* $\neg g \notin \mathrm{B}_x^{\star f}(i, \mathrm{M}, s)$ *then* $\mathrm{B}_x^{\star f+g}(i, \mathrm{M}, s) \subseteq \mathrm{B}_x^{\star f \wedge g}(i, \mathrm{M}, s)$

Proposition 5.23 provides a rephrasing of the AGM postulates for belief revision in terms of our framework. We focus on some peculiarities in this proposition, and refer for a thorough explanation on these postulates to the standard work by Gärdenfors [38]. The most remarkable cases of Proposition 5.23 are given in the fourth and the fifth item. The fourth item states that a revision with f of the x-beliefs of an agent i results in the absurd belief set iff i believes $\neg f$ with a credibility higher than x. Whenever we apply the $\texttt{revise}^x(f, i)$ functions in the interpretation of the informative actions, we will (have to) ensure that agent i does not believe $\neg f$ with a credibility higher than x, thereby preventing the agent's belief set from becoming inconsistent. The fifth item states that if agent i believes in the equivalence of f and g with a credibility higher than x, then x-revisions with f and g result in the same set of (propositional) x-beliefs.

As mentioned above, the $\texttt{revise}^o$ function implements a special kind of AGM expansion. More in particular, in states where a formula f holds, $\texttt{revise}^o(i, f)$ causes an

expansion with f of the observational beliefs of agent i, whereas in states where f does not hold, i's observational beliefs are expanded with $\neg f$.

5.24. Proposition. *For all* $\mathrm{M} \in \mathbf{M}^{\mathrm{I}}$, $s \in \mathrm{M}$, $i \in \mathrm{A}$ *and* $f \in \mathrm{L}_0$ *we have:*

- $\mathrm{M}, s \models^{\mathrm{I}} f \Rightarrow \mathrm{B}_o^{\star f}(i, \mathrm{M}, s) = \mathrm{B}_o^{+f}(i, \mathrm{M}, s)$
- $\mathrm{M}, s \models^{\mathrm{I}} \neg f \Rightarrow \mathrm{B}_o^{\star f}(i, \mathrm{M}, s) = \mathrm{B}_o^{+\neg f}(i, \mathrm{M}, s)$

5.5 Formalising observations: seeing is believing

Through observations an agent *learns whether* some proposition is true of the state in which it is residing. For artificial agents it seems to be a reasonable assumption to demand that observations are *truthful*. That is, if some observation yields information that f, then it should indeed be the case that f[3]. Observations form the most trustworthy way of acquiring information: utterances like 'I've seen it with my own eyes' or 'Seeing is believing' support this claim[4]. The formalisation that we propose is therefore such that observations overrule any beliefs acquired by other means. In situations where an agent does not observationally believe whether f, its beliefs are revised by applying the $\mathtt{revise}^o$ function to the model and the state under consideration. If the agent already observationally believed whether f no revision takes place. Note that observations will never conflict with an agent's observational beliefs or its knowledge: since both these notions are veridical, and observations are truthful, it is not possible that an observation that some formula f holds contradicts observational belief or knowledge that $\neg f$ holds, since this would force both f and $\neg f$ to be true in one and the same state.

5.25. Definition. For all $\mathrm{M} \in \mathbf{M}^{\mathrm{I}}$ with state s and all $f \in \mathrm{L}_0$ we define:

$$\mathrm{r}^{\mathrm{I}}(i, \mathtt{observe}\, f)(\mathrm{M}, s) = \begin{cases} \mathrm{M}, s & \text{if } \mathrm{M}, s \models^{\mathrm{I}} \mathbf{Bwhether}_i^o f \\ \mathtt{revise}^o(i, f)(\mathrm{M}, s), s & \text{otherwise} \end{cases}$$

The function r^{I} for actions $\mathtt{observe}\, f$ with $f \in \mathrm{L}_0$ as given in Definition 5.25 is indeed well-defined in the sense that well-formed models are transformed into well-formed models as the result of performing an observation.

[3] Note that this property does not always hold for human agents: magicians make a living out of this.

[4] There is evidence that even for artificial agents this assumption may be an oversimplification. As was pointed out by Gil Tidhar [126, 127], agents that simulate the behaviour of air combat pilots should on some occasions, i.e. when engaging in close combat, consider the observations made by their own sensors most credible, while on other occasions, i.e. when the enemy's aircrafts are beyond visual range, the information obtained from the ground controller provides for the most credible source. In situations of the latter kind, the information communicated from the ground controllers is far more credible than the one that is observed by the agent.

5.26. PROPOSITION. *For all* $\mathrm{M} \in \mathbf{M}^{\mathrm{I}}$ *with state* s*, for all* $i \in \mathrm{A}$ *and* $f \in \mathrm{L}_0$*, it holds that if* $\mathrm{M}', s = \mathtt{r}^{\mathrm{I}}(i, \mathtt{observe}\, f)(\mathrm{M}, s)$*, then* $\mathrm{M}' \in \mathrm{M}^{\mathrm{I}}_{\sim}$.

Besides being correct, Definition 5.25 is also intuitively acceptable as can be seen in the following proposition.

5.27. PROPOSITION. *For all* $i, j \in \mathrm{A}$*,* $f, g \in \mathrm{L}_0$ *and* $\varphi \in \mathrm{L}^{\mathrm{I}}$ *we have:*

1. $\mathtt{observe}\, f$ *is* o*-informative and truthful with respect to* f
2. $\mathtt{observe}\, f$ *is deterministic, idempotent and realisable*
3. $\models^{\mathrm{I}} \mathbf{K}_j g \leftrightarrow \langle \mathrm{do}_i(\mathtt{observe}\, f)\rangle \mathbf{K}_j g$
4. $\models^{\mathrm{I}} \mathbf{B}^o_j g \rightarrow \langle \mathrm{do}_i(\mathtt{observe}\, f)\rangle \mathbf{B}^o_j g$
5. $\models^{\mathrm{I}} (f \wedge \langle \mathrm{do}_i(\mathtt{observe}\, f)\rangle \mathbf{B}^o_i g) \leftrightarrow (f \wedge \mathbf{B}^o_i(f \rightarrow g))$
6. $\models^{\mathrm{I}} (\neg f \wedge \langle \mathrm{do}_i(\mathtt{observe}\, f)\rangle \mathbf{B}^o_i g) \leftrightarrow (\neg f \wedge \mathbf{B}^o_i(\neg f \rightarrow g))$
7. $\models^{\mathrm{I}} \langle \mathrm{do}_i(\mathtt{observe}\, f)\rangle \varphi \leftrightarrow \langle \mathrm{do}_i(\mathtt{observe}\, \neg f)\rangle \varphi$
8. $\models^{\mathrm{I}} f \wedge \mathbf{Ignorant}^k_i f \rightarrow \langle \mathrm{do}_i(\mathtt{observe}\, f)\rangle \mathbf{Saw}_i f$
9. $\models^{\mathrm{I}} \neg f \wedge \mathbf{Ignorant}^k_i f \rightarrow \langle \mathrm{do}_i(\mathtt{observe}\, f)\rangle \mathbf{Saw}_i \neg f$
10. $\models^{\mathrm{I}} f \wedge (\mathbf{Heard}_i \neg f \vee \mathbf{Jumped}_i \neg f) \rightarrow \langle \mathrm{do}_i(\mathtt{observe}\, f)\rangle \mathbf{Saw}_i f$
11. $\models^{\mathrm{I}} f \wedge \mathbf{B}^c_i \neg f \rightarrow \langle \mathrm{do}_i(\mathtt{observe}\, f)\rangle ((\mathbf{B}^c_i \varphi \leftrightarrow \mathbf{B}^o_i \varphi) \wedge (\mathbf{B}^d_i \varphi \leftrightarrow \mathbf{B}^o_i \varphi))$

The first item of Proposition 5.27 formalises two essential properties of observations, viz. their informativeness and truthfulness. The properties given in the second item are not uncommon for informative actions: determinism and idempotence are strongly related to the AGM postulates, and are also encountered for the other informative actions. Realisability is typical for observations; it models the idea that our agents are perfect observers which always have the opportunity to make an observation. Item 3 states that the knowledge fluents — the propositional formulae known to be true — of all agents remain unaffected under execution of an `observe` action by one of them, and item 4 states an analogous, but slightly weaker property, for their observational beliefs. The fifth and sixth item follow from the fact, formalised in Proposition 5.24, that observational beliefs are not so much revised as expanded. In item 7 it is formalised that the $\mathtt{observe}\, f$ action models 'observing whether f': observing whether f is in all aspects equivalent to observing whether $\neg f$. Items 8 and 9 state that for knowledge-ignorant agents observations actually lead to *learning by seeing*. Item 10 — a special case of item 8 — is intuitively a very nice one: it states that observations are the most credible source of information. Observations overrule other beliefs acquired through communication or adopted by default, i.e. incorrect communicational or default beliefs are *revised* in favour of observational beliefs. The last item of Proposition 5.27 sheds some more light on the (rigorous) way in which beliefs are revised: observing something that contradicts communicational beliefs leads to a *reset* of both the latter and the default beliefs of the

agent, i.e. after such a revision all the beliefs of the agent are at least grounded in its observations. Phrased differently, there no longer is any formula that the agent believes due to it being told or assuming it by default, i.e. after such a revision there is no formula ψ for which either $\mathbf{Heard}_i\psi$ or $\mathbf{Jumped}_i\psi$ holds.

5.6 Formalising communication: hearing is believing

The second source of information available to an agent consists of the information communicated by other agents. As we present it here, communication is reduced to its barest form, viz. the transfer of information. That is, we are not dealing with concepts like communication protocols, synchronisation and the like, but instead consider communication as consisting of an agent transferring some of its information to another agent. In general, agents have the opportunity to send all of their beliefs, and nothing else. Depending on the credibility of both the sending agent and the information that it sends, the receiving agent may use this information to revise its beliefs. For reasons of simplicity, we define the credibility of the sending agent as a binary notion, i.e. the agent is either credible or it is not credible, without distinguishing degrees of credibility. The notion of credibility is modelled through the so-called dependence operator, originally proposed by Huang [59]. This ternary operator, pertaining to a pair of agents and a formula, models that there is a relation of trust, dependence or credibility between the agents with respect to the formula: $\mathbf{D}_{i,j}\varphi$ indicates that agent i accepts agent j as an authority on φ, or that j is a teacher of i on the subject φ. The $\mathbf{D}_{i,j}$ operator is interpreted by incorporating a function $\mathrm{D} : \mathrm{A} \times \mathrm{A} \to \mathrm{S} \to \wp(\mathrm{L}^{\mathrm{I}})$ in the models for L^{I}: $\mathbf{D}_{i,j}\varphi$ is true in a state s of some model iff φ is in $\mathrm{D}(i,j)(s)$. The credibility of the transferred information is determined by both the credibility that the sending agent attaches to this information and the credibility that the receiving agent attaches to any of its beliefs that contradict this information. That is, if the sending agent itself observationally believes the information that it is sending, then this information overrules any contradicting beliefs that the receiving agent may have. If the sending agent itself was told the information that it is now transferring, the receiving agent will accept this information only if it does not have any contradicting information that it believes at least with the credibility attached to communicational beliefs. Information that the sending agent adopted by default is considered to be too weak to ever justify a revision of the receiving agent's beliefs. These intuitive ideas are formalised in Definition 5.28.

5.28. Definition. For all $\mathrm{M} \in \mathbf{M}^{\mathrm{I}}$ with state s, $i, j \in \mathrm{A}$, $\varphi \in \mathrm{L}^{\mathrm{I}}$ and $f \in \mathrm{L}_0$ we define:

$$\mathrm{M}, s \models^{\mathrm{I}} \mathbf{D}_{i,j}\varphi \Leftrightarrow \varphi \in \mathrm{D}(i,j)(s)$$

$$\mathtt{r}^{\mathrm{I}}(j, \mathtt{inform}\,(f,i))(\mathrm{M},s) = \begin{cases} \emptyset & \text{if } \mathrm{M},s \models^{\mathrm{I}} \neg \mathbf{B}^d_j f \\ \mathtt{revise}^c(i,f)(\mathrm{M},s), s & \text{if } \mathrm{M},s \models^{\mathrm{I}} \mathbf{D}_{i,j} f \wedge ((\mathbf{B}^o_j f \wedge \mathbf{Ignorant}^o_i f) \vee \\ & \qquad\qquad (\mathbf{Heard}_j f \wedge \mathbf{Ignorant}^c_i f)) \\ \mathrm{M}, s & \text{otherwise} \end{cases}$$

The definition of $\mathtt{r}^{\mathrm{I}}$ is also for communication actions well-defined.

5.29. Proposition. *For all* $\mathrm{M} \in \mathbf{M}^{\mathrm{I}}$ *with state* s*, for all* $i, j \in \mathrm{A}$ *and* $f \in \mathrm{L}_0$*, it holds that if* $\mathtt{r}^{\mathrm{I}}(j, \mathtt{inform}\,(f,i))(\mathrm{M},s) = \mathrm{M}', s$ *then* $\mathrm{M}' \in \mathrm{M}^{\mathrm{I}}_{\sim}$.

In the following proposition we summarise some validities describing the behaviour of the communication actions. These validities show that our formalisation indeed corresponds to the intuitive ideas unfolded above.

5.30. Proposition. *For all* $i, i', j \in \mathrm{A}$*,* $f, g \in \mathrm{L}_0$ *and* $\varphi \in \mathrm{L}^{\mathrm{I}}$*, we have:*

1. $\mathtt{inform}\,(f,i)$ *is deterministic and idempotent*
2. $\models^{\mathrm{I}} \mathbf{B}^x_{i'} g \to [\mathrm{do}_j(\mathtt{inform}\,(f,i))]\mathbf{B}^x_{i'} g$ *for* $x \in \{k, o\}$
3. $\models^{\mathrm{I}} \mathbf{B}^d_j f \leftrightarrow \langle \mathrm{do}_j(\mathtt{inform}\,(f,i))\rangle \top$
4. $\models^{\mathrm{I}} \mathbf{B}^d_j f \wedge \neg \mathbf{D}_{i,j} f \to (\langle \mathrm{do}_j(\mathtt{inform}\,(f,i))\rangle \varphi \leftrightarrow \varphi)$
5. $\models^{\mathrm{I}} \mathbf{D}_{i,j} f \wedge \mathbf{B}^c_j f \to \langle \mathrm{do}_j(\mathtt{inform}\,(f,i))\rangle \mathbf{Bwhether}^c_i f$
6. $\models^{\mathrm{I}} \mathbf{D}_{i,j} f \wedge \mathbf{B}^c_j f \wedge \mathbf{Ignorant}^c_i f \to \langle \mathrm{do}_j(\mathtt{inform}\,(f,i))\rangle \mathbf{Heard}_i f$
7. $\models^{\mathrm{I}} \mathbf{D}_{i,j} f \wedge \mathbf{Heard}_j f \wedge \mathbf{Bwhether}^c_i f \to (\langle \mathrm{do}_j(\mathtt{inform}\,(f,i))\rangle \varphi \leftrightarrow \varphi)$
8. $\models^{\mathrm{I}} \mathbf{D}_{i,j} f \wedge \mathbf{B}^c_j f \wedge \mathbf{Ignorant}^c_i f \to (\langle \mathrm{do}_j(\mathtt{inform}\,(f,i))\rangle \mathbf{B}^c_i g \leftrightarrow \mathbf{B}^c_i (f \to g))$
9. $\models^{\mathrm{I}} \mathbf{D}_{i,j} f \wedge \mathbf{B}^o_j f \wedge \mathbf{Ignorant}^o_i f \wedge \mathbf{B}^c_i \neg f \to (\langle \mathrm{do}_j(\mathtt{inform}\,(f,i))\rangle \mathbf{B}^c_i g \leftrightarrow \mathbf{B}^o_i (f \to g))$
10. $\models^{\mathrm{I}} \mathbf{D}_{i,j} f \wedge \mathbf{Jumped}_j f \to (\langle \mathrm{do}_j(\mathtt{inform}\,(f,i))\rangle \varphi \leftrightarrow \varphi)$
11. $\models^{\mathrm{I}} \mathbf{B}^x_i f \to (\langle \mathrm{do}_i(\mathtt{inform}\, f,i))\rangle \varphi \leftrightarrow \varphi)$ *for* $x \in \{k, o, c, d\}$

The first item of Proposition 5.30 states that the `inform` action also obeys the properties of determinism and idempotence that are intuitively related to the AGM postulates for belief revision. The second item states that both the knowledge and the observational belief fluents of all agents persist under execution of an `inform` action by one of them. Thus communication does exclusively affect the regions of beliefs that it should affect, viz. the communicational and default belief clusters. Note that, in contrast with the corresponding items of Propositions 5.27 and 5.36, the communicational beliefs of the receiving agent do not necessarily persist. The reason for this is given in item 9, where communicational beliefs are genuinely revised (and not just expanded). Item 3 states that agents may transfer all, and nothing but, their beliefs. Attempts to send non-beliefs are doomed to fail. Agents are therefore not even allowed to tell white lies; they are utterly honest. Note that item 3 also shows that the `inform` action is — in

contrast with the observe action — not realisable. Item 4 formalises that authority is a *conditio sine qua non* for effectual communication: if the receiving agent does not trust the sending agent on the transferred information, it lets the information pass without revising its beliefs (or changing anything else for that matter). Item 5 states that if some trustworthy agent j tells another agent i some formula f that j either knows, observed or was told, this leads to a state of affairs in which the receiving agent believes whether f at least with the credibility attached to communicational beliefs; whenever i is beforehand ignorant with regard to f on the level of communicational beliefs, the receiving agent actually *learns* f by being told (item 6). Item 7 states that if agent j tells i some formula that j itself was told while i already communicationally believes whether f, nothing changes, really. Items 8 and 9 deal with the ways in which the receiving agent's beliefs are revised as the result of it acquiring information through communication. If the sending agent j communicationally believes the transferred formula f and the receiving agent i is communicationally ignorant with respect to f, then an expansion with f of the communicational beliefs of i takes place (item 8). If j observationally believes f and i is observationally ignorant with respect to f while communicationally believing $\neg f$, then i's communicational beliefs consist henceforth of the observational beliefs that are implied by f (item 9). Item 10 states that default beliefs are not transferable: the credibility of this kind of information is too low for it to have any effect upon being heard. The last item shows that agents cannot increase the credibility they attach to information by talking to themselves, since this talking to themselves does not change anything. Thus our agents are not susceptible to this kind of autosuggestion.

5.7 Formalising default jumps: jumping is believing

The last possible source of information that we consider in this chapter is the only endogenous one, and consists of the possibility to adopt beliefs by default. In general in default reasoning, plausible yet fallible conclusions are derived on the basis of the presence of certain information and the absence of other information. In the formalisation of default reasoning as proposed by Reiter [112], defaults are formalised as special inference rules $\varphi_1 : \varphi_2/\varphi_3$, which should be interpreted as stating that if φ_1 holds and it is consistent to assume φ_2 then φ_3 may be derived. Here we consider the most basic form of default reasoning, in which no information is required to be present and the information that needs to be absent is strongly related to the conclusion that is to be derived. In terms of Reiter's framework, the kind of default reasoning that we consider here uses only *supernormal* defaults, i.e. defaults of the form $: \varphi/\varphi$; these defaults can be seen as *possible hypotheses* in Poole's system [103]. Instead of introducing these supernormal defaults as an additional syntactical construct, we want to formalise them by using concepts already present in our language. To this end we need a modal (epistemic)

translation of defaults. If one looks at the modal translations of defaults that have been proposed in the literature [91], it turns out that for supernormal defaults $:\varphi/\varphi$ in an epistemic S5 framework all of these translations amount to either the formula $\mathbf{M}_i\varphi \rightarrow \varphi$ or $\mathbf{M}_i\varphi \rightarrow \mathbf{K}_i\varphi$. These translations stem from the usual, *static*, account of default reasoning and are therefore not completely suitable for our *dynamic* framework, where default reasoning is formalised by informative actions. Intuitively, our notion of defaults corresponds to the antecedent of both the implications given above, whereas the consequents of these implications correspond to possible results of attempted jumps. Therefore it seems reasonable to consider the formula $\mathbf{M}_i\varphi$ as a candidate to represent our kind of defaults. However, in our opinion this formalisation would do no justice to the empirical character of defaults. More in particular, the idea of defaults being rooted in *common* sense is not visible when formalising defaults as ordinary epistemic possibilities. In our multi-agent system, *common* sense is related to the knowledge and lack of knowledge of *all* agents. To capture this idea of defaults as determined by the (lack of) knowledge of all agents, we propose the modality of *common possibility*, intuitively corresponding to *being considered epistemically possible by all agents*. Although at first sight defaults as common possibilities have an *optimistic* flavour to them, agents that jump to these formulae are not that bold at all: there is nothing in the joint knowledge of all agents that could deem these jumps inappropriate.

5.31. Definition. For $\varphi \in \mathrm{L}^{\mathrm{I}}$, the formula $\mathbf{N}\varphi$, for nobody knows not φ, is defined by:

$$\mathbf{N}\varphi \triangleq \mathbf{M}_1\varphi \wedge \ldots \wedge \mathbf{M}_n\varphi$$

5.32. Remark. In order for the common possibility operator to be well-defined, the set of agents has to be finite, since the languages considered in this thesis do not allow for infinite conjunctions. It is furthermore obvious that, should the common possibility operator express a notion different from ordinary, epistemic possibility, the set of agents should contain at least two elements. These arguments provide (additional) support for the constraints imposed in Definition 5.1.

Not surprisingly, the common possibility operator $\mathbf{N}$ shares some of the properties of the epistemic possibility operator $\mathbf{M}_i$. In particular, $\mathbf{N}$ satisfies the dual KT4-axiomatisation, but satisfies only one direction of the dual 5-axiom. Moreover, whereas the epistemic possibility operator satisfies the axiom of weak belief ($\mathbf{M}_i(\varphi \vee \psi) \leftrightarrow (\mathbf{M}_i\varphi \vee \mathbf{M}_i\psi)$), the common possibility operator does not satisfy the left-to-right implication.

5.33. Proposition. *For all $\varphi, \psi \in \mathrm{L}^{\mathrm{I}}$ we have:*

1. $\models^{\mathrm{I}} \varphi \rightarrow \mathbf{N}\varphi$
2. $\models^{\mathrm{I}} \mathbf{N}\varphi \vee \mathbf{N}\neg\varphi$

3. $\models^{\mathrm{I}} \mathbf{N}\mathbf{N}\varphi \to \mathbf{N}\varphi$
4. $\models^{\mathrm{I}} \neg\mathbf{N}\varphi \to \mathbf{N}\neg\mathbf{N}\varphi$
5. $\mathbf{N}\neg\mathbf{N}\varphi \to \neg\mathbf{N}\varphi$ *is not for all* $\varphi \in \mathrm{L}^{\mathrm{I}}$ *valid*
6. $\mathbf{N}(\varphi \vee \psi) \to \mathbf{N}\varphi \vee \mathbf{N}\psi$ *is not for all* $\varphi, \psi \in \mathrm{L}^{\mathrm{I}}$ *valid*
7. $\models^{\mathrm{I}} \mathbf{N}\varphi \to \mathbf{N}(\varphi \vee \psi)$
8. $\models^{\mathrm{I}} \varphi \to \psi \Rightarrow \models^{\mathrm{I}} \mathbf{N}\varphi \to \mathbf{N}\psi$
9. $\mathbf{N}\varphi \wedge \mathbf{N}(\varphi \to \psi) \to \mathbf{N}\psi$ *is not for all* $\varphi, \psi \in \mathrm{L}^{\mathrm{I}}$ *valid*
10. $\mathbf{N}\varphi \wedge \mathbf{N}\psi \to \mathbf{N}(\varphi \wedge \psi)$ *is not for all* $\varphi, \psi \in \mathrm{L}^{\mathrm{I}}$ *valid*
11. $\mathbf{N}\varphi \to \neg\mathbf{N}\neg\varphi$ *is not for all* $\varphi \in \mathrm{L}^{\mathrm{I}}$ *valid*

The properties formalised in Proposition 5.33 indicate that $\mathbf{N}$ is in most respects indeed a possible candidate to represent defaults. Item 1 states that true formulae are defaults. This is a consequence of the reflexivity of the epistemic accessibility relation that ensures veridicality of knowledge. For each agent considers all formulae that hold in the 'current' state to be epistemically possible, and hence all these formulae are common possibilities. Item 2 indicates that in principle all gaps in the agent's information are 'fillable' by a default belief, i.e. for all formulae φ either φ or $\neg\varphi$ is a default. Possibly the most important and remarkable property of the common possibility operator with regard to its usability to represent defaults is given by item 6, which indicates that disjunctive defaults are not necessarily trivialised, i.e. these disjunctions are not necessarily reduced to their disjuncts. This property is very important for the expressive power of our framework. Consider for instance the situation of a lottery with $m \gg 1$ players (cf. [75]). Then it is not the case that player 1 wins by default, and neither is this the case for any of the players 2 to $m \Leftrightarrow 1$. But it is also the case that by default one of these $m \Leftrightarrow 1$ players actually does win. Since $\{\neg\mathbf{N}w_1, \ldots, \neg\mathbf{N}w_{m-1}, \mathbf{N}(w_1 \vee \ldots \vee w_{m-1})\}$ is satisfiable, this aspect of the lottery can be formalised in our framework. Note that this situation cannot be formalised by taking ordinary (single-agent) epistemic possibility instead of common possibility, since $\mathbf{M}_i(\varphi \vee \psi) \leftrightarrow \mathbf{M}_i\varphi \vee \mathbf{M}_i\psi$ is a valid formula. Items 10 and 11 show that the Nixon-diamond (cf. [113]) can be represented. That is, it is possible to represent that it is a default that Nixon was a pacifist and that it is a default that he was a non-pacifist, even though it is not a default that he was a walking contradiction.

Using the common possibility operator to represent defaults, we now come to the formalisation of the attempted jumps to conclusion that constitute (supernormal) default reasoning. Since adopting a belief by default accounts for acquiring information of the lowest credibility, it is obvious that default jumps are effective only for agents that are completely ignorant with respect to the default that is jumped to.

5.34. Definition. For all $\mathrm{M} \in \mathbf{M}^{\mathrm{I}}$ with state s, $i \in \mathrm{A}$, and $f \in \mathrm{L}_0$ we define:

$\mathrm{r}^{\mathrm{I}}(i, \mathtt{try\_jump}\, f)(\mathrm{M}, s) =$

$$\begin{cases} \emptyset & \text{if } \mathrm{M}, s \models^{\mathrm{I}} \neg \mathbf{N} f \\ \mathtt{revise}^d(i, f)(\mathrm{M}, s), s & \text{if } \mathrm{M}, s \models^{\mathrm{I}} \mathbf{N} f \wedge \mathbf{Ignorant}_i^d f \\ \mathrm{M}, s & \text{otherwise} \end{cases}$$

Also for this last kind of informative actions, the function r^{I} is well-defined.

5.35. Proposition. *For all* $\mathrm{M} \in \mathbf{M}^{\mathrm{I}}$ *with state* s*, for all* $i \in \mathrm{A}$ *and* $f \in \mathrm{L}_0$*, it holds that if* $\mathrm{r}^{\mathrm{I}}(j, \mathtt{try\_jump}\, f)(\mathrm{M}, s) = \mathrm{M}', s$ *then* $\mathrm{M}' \in \mathrm{M}^{\mathrm{I}}_{\sim}$.

In addition to being correct, Definition 5.34 is also intuitively acceptable, which is shown in Proposition 5.36.

5.36. Proposition. *For all* $i, j \in \mathrm{A}$*,* $f, g \in \mathrm{L}_0$ *and* $\varphi \in \mathrm{L}^{\mathrm{I}}$*, we have:*

1. $\mathtt{try\_jump}\, f$ *is deterministic, idempotent and* d*-informative with regard to* f
2. $\models^{\mathrm{I}} \mathbf{B}_j^x g \rightarrow [\mathrm{do}_i(\mathtt{try\_jump}\, f)]\mathbf{B}_j^x g$ *for* $x \in \{d, c, o, k\}$
3. $\models^{\mathrm{I}} \mathbf{N} f \leftrightarrow \langle \mathrm{do}_i(\mathtt{try\_jump}\, f)\rangle \top$
4. $\models^{\mathrm{I}} \langle \mathrm{do}_i(\mathtt{try\_jump}\, f)\rangle \top \leftrightarrow \langle \mathrm{do}_i(\mathtt{try\_jump}\, f)\rangle \mathbf{Bwhether}_i^d f$
5. $\models^{\mathrm{I}} \mathbf{N} f \wedge \mathbf{Ignorant}_i^d f \rightarrow \langle \mathrm{do}_i(\mathtt{try\_jump}\, f)\rangle \mathbf{Jumped}_i f$
6. $\models^{\mathrm{I}} \mathbf{N} f \wedge \mathbf{Ignorant}_i^d f \rightarrow (\langle \mathrm{do}_i(\mathtt{try\_jump}\, f)\rangle \mathbf{B}_i^d g \leftrightarrow \mathbf{B}_i^d (f \rightarrow g))$
7. $\models^{\mathrm{I}} \mathbf{N} f \wedge \mathbf{Bwhether}_i^d f \rightarrow (\langle \mathrm{do}_i(\mathtt{try\_jump}\, f)\rangle \varphi \leftrightarrow \varphi)$

The first item of Proposition 5.36 deals again with the properties more or less typical for informative actions. It is obvious that the property of realisability is not validated since an agent may attempt to jump to a non-default, and in 3 it is formalised that such attempted jumps are doomed to fail. Item 2 states that all information of all agents persists under the attempted jump to a formula by one of them. Item 4 states that default jumps for which an agent has the opportunity always result in the agent believing by default whether the formula that is jumped to holds. The fifth item formalises the idea that agents that are completely ignorant with respect to some formula f jump to new default beliefs by applying the $\mathtt{try\_jump}\, f$ action. The incorporation of these new beliefs is brought about by expanding the default beliefs of the agent with the default that is adopted (item 6). The last item states that attempted jumps to default conclusions yield information for totally ignorant agents only.

5.8 The ability to gather information

As we see it, in the ability of intelligent information agents to execute informative actions two different notions are combined. On the one hand, the abilities of an agent restrict its practical possibility to acquire information: only the actions that are within the agent's capacities can be used as means to extend its information. This corresponds to the idea

that agents are not just able to acquire all the information they would like to acquire. On the other hand, we use the agents' abilities to steer the way in which information is acquired. That is, through its abilities an agent is forced to prefer more credible sources of information. We will, for instance, define an agent to be able to try to adopt some formula by default only if it cannot acquire this information through its — more credible — observations.

Since, by nature, observations provide the most credible source of information, the ability to observe is determined strictly by the fact that the agents' information gathering is limited, and does not depend on other means to acquire information. In our opinion it is reasonable to consider this limit to the agent's ability to perform observations as given by the construction of the agent. For instance, human agents are built in such a way that they cannot observe objects at great distances, or objects that are outside the spectrum of human observation, like for instance X-rays. In the case of artificial agents, one could think of two robots, working together to explore a strange, new world. Each robot is equipped with its own, personal set of sensors: one robot is for instance able to observe whether its environment is radioactive, the other whether the planet's atmosphere contains oxygen, but neither robot is able to observe both. Being a construction decision, we assume that the observational capacities of agents are determined beforehand, i.e. with respect to the agents' ability observations are treated as being atomic.

In defining the capabilities of agents to inform other agents, we consider the *moral* component of ability to be most relevant. That is, we demand our agents to be sincere in that they are morally unable to lie or gossip. The things that an agent is capable of telling to other agents are exactly the things that it itself believes, be it with the lowest credibility. In this way the information acquisition of an agent that is due to communication is restricted to those formulae that are believed by some authority.

The ability to attempt to jump to a default captures both aspects described above, i.e. both the aspect of restricted information acquisition as well as that of preferring the most credible source of information are visible in the definition of the ability to (attempt to) jump. Concerning the first aspect, an agent is able to jump to a default only if it *knows* it to be a default, i.e. agents have to know their defaults in order to be able to use them. With respect to the second aspect, the capability to jump is defined to depend on the observational capacities of the agent: an agent is able to attempt a jump to a formula only if it knows that it is not able to observe whether the formula holds. In this way it is ensured that agents resort to default jumps only if the possibility of acquiring the information through observations is excluded.

5.37. Definition. Let $\mathrm{M} \in \mathbf{M}^{\mathrm{I}}$ be some model. The function c^{I} is for the informative actions defined as follows, where s is some state in M, i, j are agents and $f \in \mathrm{L}_0$ is some

propositional formula.

$$\begin{aligned} \mathrm{c}^{\mathrm{I}}(j, \mathtt{inform}\,(f,i))(\mathrm{M},s) &= \mathrm{c}^{\mathrm{I}}(j, \mathtt{confirm}\,\mathbf{B}^d_j f)(\mathrm{M},s) \\ \mathrm{c}^{\mathrm{I}}(i, \mathtt{try\_jump}\,f)(\mathrm{M},s) &= \mathrm{c}^{\mathrm{I}}(i, \mathtt{confirm}\,\mathbf{K}_i(\neg\mathbf{A}_i\mathtt{observe}\,f \wedge \mathbf{N}f))(\mathrm{M},s) \end{aligned}$$

where $\lambda f \in \mathrm{L}_0.(\mathrm{c}^{\mathrm{I}}(i, \mathtt{observe}\,f)(\mathrm{M},s))$ is a function in f such that

$$\mathrm{c}^{\mathrm{I}}(i, \mathtt{observe}\,f)(\mathrm{M},s) = \mathrm{c}^{\mathrm{I}}(i, \mathtt{observe}\,\neg f)(\mathrm{M},s)$$

5.38. Proposition. *For all* $i, j \in \mathrm{A}$ *and* $f \in \mathrm{L}_0$ *we have:*

1. $\models^{\mathrm{I}} \mathbf{A}_i\mathtt{observe}\,f \leftrightarrow \mathbf{A}_i\mathtt{observe}\,\neg f$
2. $\models^{\mathrm{I}} \mathbf{A}_j\mathtt{inform}\,(f,i) \leftrightarrow \langle \mathrm{do}_j(\mathtt{inform}\,(f,i))\rangle\top$
3. $\models^{\mathrm{I}} \mathbf{A}_i\mathtt{try\_jump}\,f \rightarrow \langle \mathrm{do}_i(\mathtt{try\_jump}\,f)\rangle\top$
4. $\models^{\mathrm{I}} \mathbf{A}_i\mathtt{observe}\,f \rightarrow \neg\mathbf{A}_i\mathtt{try\_jump}\,f$

The first item of Proposition 5.38 again expresses that the observe actions formalise 'observing whether'. Item 2 and 3 state that both the actions modelling communication and those modelling default reasoning, though not realisable *per se*, are *A-realisable*, i.e. having the ability to perform these actions implies having the opportunity to do so. As mentioned in Chapter 2, the property of A-realisability is considered unacceptable for mundane actions, but given our view on the notion of ability for the non-mundane, informative actions this property seems much less controversial. For both the ability and the opportunity to perform informative actions are defined in terms of informational attitudes, and thus the distinction between these notions is less clear than it is for mundane actions. The last item of Proposition 5.38 expresses that agents are able to attempt to jump to a formula only if it is not within their capacities to observe whether the formula holds.

5.9 Summary and conclusions

In this chapter we presented a formal framework that may be seen as a step towards a formalisation of intelligent information agents. This formalisation concerns the agents' informational attitudes, the actions that the agents may perform, and in particular the interaction between actions and information. In order to solve information conflicts, we imposed a credibility ordering both on the beliefs of an agent and on its possible sources of information. This ordering is defined by dividing the information of an agent into four sets, situated within each other. The credibility of an item of information is determined by the smallest set that it is an element of. To model the act of acquiring information, we introduced special, so-called informative actions. By performing an informative action, an agent may acquire (additional) information. We considered three

kinds of informative actions, which correspond to information acquisition by observation, communication and default, respectively. Of these, observations provide the information with the highest credibility, while communication is in general considered to be a more credible source than reasoning by default. The changes in the beliefs of an agent that result from the execution of an informative action are brought about by changing the model under consideration, which accounts for a model-transforming interpretation of informative actions rather than the standard state-transition one. The transformation of models is such that the beliefs of an agent are changed in compliance with the AGM postulates for belief revision. Through the abilities of agents we defined both the limited capacities of agents to acquire information and the priority that agents should give to the most credible information source. On the whole, the formal framework can in our opinion rightfully be seen as a formalisation of intelligent information agents.

5.9.1 Possible extensions

One can think of several obvious yet interesting extensions of the framework presented in this chapter, most of which are in fact not too hard to implement. The first of these would be to consider a more refined way of revising the agents' beliefs, i.e. instead of the straightforward AiG revision one could think of using a more refined and intuitively better revision. By using the machinery presented in [81, 84] we could not only implement a more refined revision, but also formally account for iterated revisions of belief. The second possible extension was suggested by Castelfranchi [15], and consists of refining the communication part of information acquisition. In this more refined form of communication it should be possible to relate the credibility that an agent attaches to some formula that it has been told, to the number of agents that have been communicating this formula, i.e. the more agents that tell that some formula is true, the more fiercely it is believed. It is possible to formalise this intuitive idea by splitting the communicational belief cluster of an agent into different parts, one for each of the other agents. Whenever an agent j informs an agent i of the truth of f, i's communicational belief cluster that is associated with j is revised with f. Over the different communicational belief clusters a kind of *graded* belief modality could be defined, which formalises the credibility attached to the agent's communicational beliefs. The formula $\mathbf{B}_i^{c,0.5} f$ would then be taken to represent that agent i communicationally believes f with credibility 0.5, i.e. f holds in at least half of the communicational belief clusters of i. It is clear that in this way it can indeed be modelled that the credibility attached by i to one of its communicational beliefs depends on the number of agents that have informed i of the truth of this formula. Another possibility to refine communication could be to allow a form of 'demand-driven' communication in addition to the 'supply-driven' one considered here. In this extended form of communication an agent may request information on

the truth or falsity of certain proposition from another, trusted, agent. One possible way of formalising this kind of extended communication is given in [82], where an additional modal operator is used to record the agents' requests. Yet another possible extension that we would like to mention here concerns the possibility of agents to reason by default. Instead of restricting oneself to supernormal defaults, one could consider more general defaults. Rather than using (epistemic) concepts already present in the framework, these defaults would probably have to be modelled by some additional means (like a function yielding the defaults available to an agent in a state), but as soon as these means are provided, there is nothing to prevent one from using more general defaults.

Another very interesting extension of the framework presented in this chapter concerns the incorporation of actions associated with belief *updates* (cf. [67]) in addition to the ones associated with belief revision that we considered here. Whereas belief revisions are changes in information on an unchanging world, updates are information changes that are associated with a change in the state of the world. The interaction between these different information changes might well be worth looking at.

The framework could also be extended to make it a suitable formalisation tool for special agents, like *intelligent information retrieval agent*. These agents assist a user with finding relevant information, particularly in cyberspace, that satisfies one of his/her information needs. To this end they communicate, either with their user or with other (intelligent information retrieval) agents, go to World Wide Web sites to seek for relevant information, or make assumptions by default. A preliminary formalisation of these intelligent information retrieval agents based on the framework proposed in this chapter was presented by Huibers & Van Linder [62, 63].

5.9.2 Bibliographical notes

The backbone of this chapter is an extended and revised version of [87], in which elements of various other papers have been incorporated. In Sections 5.3 and 5.4, parts of [84] have been used, Section 5.5 uses [83], in Section 5.6 elements of [82] were used and in Section 5.7 we used [85].

There is not much related work on the formalisation of (aspects of) intelligent information agents that we know of. Kraus & Lehmann were the first to combine knowledge, belief and time [71], but the only genuinely relevant related work that we know of is the formalisation of *knowledge-producing actions* in the Situation Calculus as proposed by Scherl & Levesque [116]. The term 'knowledge-producing' as used by Scherl & Levesque coincides for the greater part with our use of the term 'informative'. The most important characteristics of knowledge-producing actions is that upon execution knowledge is produced while leaving all other fluents unaffected. To formalise knowledge, Scherl & Levesque adapt the standard modal possible worlds model to the Situation Calculus:

instead of an epistemic accessibility relation as used in our framework, a relation over situations is used. The interpretation of knowledge-producing actions resembles our interpretation of informative actions to some extent. When performing an action $\texttt{Sense}\,P$, which represents the act of sensing whether P holds, in a situation S, a situation S' results such that the only situations considered epistemically possible in S' agree with S' on the truth-value assigned to P, while causing as little change as possible. Since only one informational attitude is considered, viz. knowledge, which is veridical, and knowledge-producing actions are by definition truthful, there is no question of AGM-like revision procedures as is the case in our framework. That is, knowledge-producing actions are really *knowledge-producing*, and not knowledge-revising. Still, in spite of this difference and the ones due to the conceptual and ontological distinctions between the framework underlying the formalisation of Scherl & Levesque — the Situation Calculus — and that underlying ours, the treatment of informative and that of knowledge-producing actions are surprisingly similar.

5.10 Selected proofs

5.8. Proposition. *For all i-doxastic sequenced formulae $\chi \in \mathrm{L}^{\mathrm{I}}$, for all $\varphi \in \mathrm{L}^{\mathrm{I}}$ and for all $\mathbf{X} \in \{\mathbf{B}_i^o, \mathbf{B}_i^c, \mathbf{B}_i^d\}$ we have:*

1. $\models^{\mathrm{I}} \mathbf{X}\chi \leftrightarrow \chi$
2. $\models^{\mathrm{I}} \mathbf{X}\mathbf{K}_i\varphi \leftrightarrow \mathbf{K}_i\varphi$
3. $\models^{\mathrm{I}} \mathbf{X}\neg\mathbf{K}_i\varphi \leftrightarrow \neg\mathbf{K}_i\varphi$
4. $\mathbf{X}\varphi \rightarrow \mathbf{K}_i\mathbf{X}\varphi$ *is not for all* $\varphi \in \mathrm{L}^{\mathrm{I}}$ *valid*
5. $\neg\mathbf{X}\varphi \rightarrow \mathbf{K}_i\neg\mathbf{X}\varphi$ *is not for all* $\varphi \in \mathrm{L}^{\mathrm{I}}$ *valid*

Proof: We show the first, fourth and fifth item; the second and third are obvious. So let χ be an i-doxastic sequenced formula $\mathbf{Y}\rho$, and let M be some model with state s. Slightly abusing notation we define $\mathrm{Y}(i,s)$ to be $\mathrm{Y}(i,s)$ for $\mathrm{Y} = \mathrm{B}^c, \mathrm{B}^d$ and $[s]_{\mathrm{Y}(i)}$ for $\mathrm{Y} = \mathrm{B}^o$. To prove the first item, we distinguish two cases:

- $\mathbf{Y} \in \{\mathbf{B}_i^o, \mathbf{B}_i^c, \mathbf{B}_i^d\}$. In this case we have:
  $$\begin{aligned}
  &\mathrm{M}, s \models^{\mathrm{I}} \chi \\
  \Leftrightarrow\ &\mathrm{M}, s \models^{\mathrm{I}} \mathbf{Y}\rho \\
  \Leftrightarrow\ &\forall s' \in \mathrm{Y}(i,s)(\mathrm{M}, s' \models^{\mathrm{I}} \rho) \\
  \overset{*}{\Leftrightarrow}\ &\forall s'' \in \mathrm{X}(i,s)\forall s' \in \mathrm{Y}(i,s'')(\mathrm{M}, s' \models^{\mathrm{I}} \rho) \\
  \Leftrightarrow\ &\forall s'' \in \mathrm{X}(i,s)(\mathrm{M}, s'' \models^{\mathrm{I}} \mathbf{Y}\rho) \\
  \Leftrightarrow\ &\mathrm{M}, s \models^{\mathrm{I}} \mathbf{X}\mathbf{Y}\rho \\
  \Leftrightarrow\ &\mathrm{M}, s \models^{\mathrm{I}} \mathbf{X}\chi
  \end{aligned}$$
- $\mathbf{Y} \in \{\neg\mathbf{B}_i^o, \neg\mathbf{B}_i^c, \neg\mathbf{B}_i^d\}$. In this case we have:

$$
\begin{aligned}
& \mathrm{M}, s \models^{\mathrm{I}} \chi \\
\Leftrightarrow\ & \mathrm{M}, s \models^{\mathrm{I}} \mathbf{Y}\rho \\
\Leftrightarrow\ & \exists s' \in \mathrm{Y}(i,s)(\mathrm{M}, s' \models^{\mathrm{I}} \neg\rho) \\
\overset{\star}{\Leftrightarrow}\ & \forall s'' \in \mathrm{X}(i,s)\exists s' \in \mathrm{Y}(i,s'')(\mathrm{M}, s' \models^{\mathrm{I}} \neg\rho) \\
\Leftrightarrow\ & \forall s'' \in \mathrm{X}(i,s)(\mathrm{M}, s'' \models^{\mathrm{I}} \mathbf{Y}\rho) \\
\Leftrightarrow\ & \mathrm{M}, s \models^{\mathrm{I}} \mathbf{X}\mathbf{Y}\rho \\
\Leftrightarrow\ & \mathrm{M}, s \models^{\mathrm{I}} \mathbf{X}\chi
\end{aligned}
$$

The equivalences tagged with $\star$ hold since for $\mathrm{Y} \in \{\mathrm{B}^o, \mathrm{B}^c, \mathrm{B}^d\}$, $\mathrm{Y}(i,s) \neq \emptyset$, $\mathrm{Y}(i,s) \subseteq [s]_{\mathrm{B}^o(i)}$ and $\mathrm{Y}(i,s'') = \mathrm{Y}(i,s)$ for all $s'' \in [s]_{\mathrm{B}^o(i)}$. Since the two items given above capture all possible i-doxastic sequenced formulae χ we conclude that the first item of Proposition 5.8 holds.

To show the fourth and the fifth item, consider the model M which is for some $i \in \mathrm{A}$ and $p \in \Pi$ such that $\mathrm{S} = \{s_0, s_1\}$, $\pi(p, s_0) = \mathbf{1}$ and $\pi(p, s_1) = \mathbf{0}$, $\mathrm{R}(i) = \mathrm{S}^2$, $\mathrm{B}^o(i) = \{(s_l, s_l)\}, \mathrm{B}^c(i, s_l) = \mathrm{B}^d(i, s_l) = \{s_l\}$ for $l = 0, 1$, and $\mathtt{r}_0$ and $\mathtt{c}_0$ are arbitrary. For this model it holds that $\mathrm{M}, s_0 \models^{\mathrm{I}} \mathbf{X}p \wedge \neg\mathbf{K}_i\mathbf{X}p$ and $\mathrm{M}, s_1 \models^{\mathrm{I}} \neg\mathbf{X}p \wedge \neg\mathbf{K}_i\neg\mathbf{X}p$ for $\mathbf{X} \in \{\mathbf{B}_i^o, \mathbf{B}_i^c, \mathbf{B}_i^d\}$, which suffices to conclude items four and five.
⊠

5.12. Proposition. *Let $\varphi, \psi \in \mathrm{L}^{\mathrm{I}}$ and $i \in \mathrm{A}$ be arbitrary. Let $\mathbf{X}$ be in the set $\{\mathbf{Heard}_i, \mathbf{Jumped}_i\}$, $\mathbf{Y} \in Bel = \{\mathbf{Saw}_i, \mathbf{Heard}_i, \mathbf{Jumped}_i\}$, and let $\mathbf{Z} \in Bel \cup \{\mathbf{B}_i^k\}$. Define the ordering $\geq'$ to be the reflexive and transitive closure of $>'$ with $\mathbf{B}_i^k >' \mathbf{Saw}_i >' \mathbf{Heard}_i >' \mathbf{Jumped}_i$. Then we have:*

1.	$\models^{\mathrm{I}} \mathbf{Y}\varphi \wedge \mathbf{Y}(\varphi \to \psi) \to \mathbf{Y}\psi$	*K*
2.	$\models^{\mathrm{I}} \neg(\mathbf{Y}\varphi \wedge \mathbf{Y}\neg\varphi)$	*D*
3.	$\models^{\mathrm{I}} \mathbf{Saw}_i\varphi \to \varphi$	*T*
4.	$\models^{\mathrm{I}} \mathbf{Y}\varphi \to \mathbf{B}_i^o\mathbf{Y}\varphi$	*4*
5.	$\models^{\mathrm{I}} \mathbf{X}\varphi \to \neg\mathbf{X}\mathbf{X}\varphi$	*4'*
6.	$\models^{\mathrm{I}} \mathbf{Saw}_i\varphi \wedge \mathbf{M}_i\neg\mathbf{Saw}_i\varphi \leftrightarrow \mathbf{Saw}_i\mathbf{Saw}_i\varphi$	*4"*
7.	$\models^{\mathrm{I}} \neg\mathbf{Y}\varphi \to \mathbf{B}_i^o\neg\mathbf{Y}\varphi$	*5*
8.	$\models^{\mathrm{I}} \neg\mathbf{X}\varphi \to \neg\mathbf{X}\neg\mathbf{X}\varphi$	*5'*
9.	$\models^{\mathrm{I}} \neg\mathbf{Saw}_i\varphi \wedge \mathbf{M}_i\mathbf{Saw}_i\varphi \leftrightarrow \mathbf{Saw}_i\neg\mathbf{Saw}_i\varphi$	*5"*
10.	$\models^{\mathrm{I}} (\mathbf{Y}\varphi \wedge \mathbf{Y}\psi) \to \mathbf{Y}(\varphi \wedge \psi)$	*C*
11.	$\mathbf{Y}(\varphi \wedge \psi) \to (\mathbf{Y}\varphi \wedge \mathbf{Y}\psi)$ *is not for all* $\varphi, \psi \in \mathrm{L}^{\mathrm{I}}$ *valid*	*M*
12.	$\models^{\mathrm{I}} \mathbf{Y}(\varphi \wedge \psi) \to (\bigvee_{\mathbf{Z} \geq' \mathbf{Y}} \mathbf{Z}\varphi \wedge \bigvee_{\mathbf{Z} \geq' \mathbf{Y}} \mathbf{Z}\psi) \wedge (\mathbf{Y}\varphi \vee \mathbf{Y}\psi)$	*M'*
13.	$\models^{\mathrm{I}} \varphi \Rightarrow \models^{\mathrm{I}} \neg\mathbf{Y}\varphi$	*N*
14.	*not for all* $\varphi, \psi \in \mathrm{L}^{\mathrm{I}}$ *does* $\models^{\mathrm{I}} \varphi \to \psi$ *imply* $\models^{\mathrm{I}} \mathbf{Y}\varphi \to \mathbf{Y}\psi$	*RM*
15.	$\models^{\mathrm{I}} \varphi \to \psi \Rightarrow \models^{\mathrm{I}} \mathbf{Y}\varphi \to \bigvee_{\mathbf{Z} \geq' \mathbf{Y}} \mathbf{Z}\psi$	*RM'*
16.	$\models^{\mathrm{I}} \varphi \leftrightarrow \psi \Rightarrow \models^{\mathrm{I}} \mathbf{Y}\varphi \leftrightarrow \mathbf{Y}\psi$	*RE*

PROOF: For each case, the proofs for the different operators are highly analogous. Therefore, we mostly restrict ourselves to proving the case for the $\mathbf{Heard}_i$ operator. At various places in the proofs we use the equivalent formulations of the derived belief operators that are proposed in Remark 5.11. So let $\mathrm{M} \in \mathbf{M}^{\mathrm{I}}$ with state s, $i \in \mathrm{A}$, and $\varphi, \psi \in \mathrm{L}^{\mathrm{I}}$ be arbitrary.

1. Suppose $\mathrm{M}, s \models^{\mathrm{I}} \mathbf{Heard}_i\varphi \wedge \mathbf{Heard}_i(\varphi \to \psi)$, i.e. $\mathrm{M}, s \models^{\mathrm{I}} \mathbf{B}_i^c\varphi \wedge \neg\mathbf{B}_i^o\varphi \wedge \mathbf{B}_i^c(\varphi \to \psi) \wedge \neg\mathbf{B}_i^o(\varphi \to \psi)$. Then, since $\mathbf{B}_i^c$ validates the K-axiom, it holds that $\mathrm{M}, s \models^{\mathrm{I}} \mathbf{B}_i^c\psi$, hence we have to show that $\mathrm{M}, s \not\models^{\mathrm{I}} \mathbf{B}_i^o\psi$. Assume towards a contradiction that $\mathrm{M}, s \models^{\mathrm{I}} \mathbf{B}_i^o\psi$. Then also $\mathrm{M}, s \models^{\mathrm{I}} \mathbf{B}_i^o(\varphi \to \psi)$ since $\mathbf{B}_i^o$ validates the K-axiom. Since this contradicts the fact that $\mathrm{M}, s \models^{\mathrm{I}} \neg\mathbf{B}_i^o(\varphi \to \psi)$, we conclude that $\mathrm{M}, s \not\models^{\mathrm{I}} \mathbf{B}_i^o\psi$. Hence $\mathrm{M}, s \models^{\mathrm{I}} \mathbf{B}_i^c\psi \wedge \neg\mathbf{B}_i^o\psi$ and thus $\mathrm{M}, s \models^{\mathrm{I}} \mathbf{Heard}_i\psi$.
2. Since $\mathbf{B}_i^c$ satisfies the D-axiom, it holds that $\mathrm{M}, s \models^{\mathrm{I}} \neg(\mathbf{B}_i^c\varphi \wedge \mathbf{B}_i^c\neg\varphi)$. Hence also $\mathrm{M}, s \models^{\mathrm{I}} \neg(\mathbf{Heard}_i\varphi \wedge \mathbf{Heard}_i\neg\varphi)$.
3. If $\mathrm{M}, s \models^{\mathrm{I}} \mathbf{Saw}_i\varphi$ this implies that $\mathrm{M}, s \models^{\mathrm{I}} \mathbf{B}_i^o\varphi$. Since $\mathbf{B}_i^o$ validates the T-axiom it follows that $\mathrm{M}, s \models^{\mathrm{I}} \varphi$.
4. This item is a direct consequence of Corollary 5.9. For $\mathrm{M}, s \models^{\mathrm{I}} \mathbf{Heard}_i\varphi$ iff $\mathrm{M}, s \models^{\mathrm{I}} \mathbf{B}_i^c\varphi \wedge \neg\mathbf{B}_i^o\varphi$, which, by Corollary 5.9, is equivalent to $\mathrm{M}, s \models^{\mathrm{I}} \mathbf{B}_i^o\mathbf{B}_i^c\varphi \wedge \mathbf{B}_i^o\neg\mathbf{B}_i^o\varphi$, and thus to $\mathrm{M}, s \models^{\mathrm{I}} \mathbf{B}_i^o\mathbf{Heard}_i\varphi$.
5. This item follows directly from the previous one: if $\mathrm{M}, s \models^{\mathrm{I}} \mathbf{Heard}_i\varphi$ then $\mathrm{M}, s \models^{\mathrm{I}} \mathbf{B}_i^o\mathbf{Heard}_i\varphi$, hence $\mathrm{M}, s \models^{\mathrm{I}} \neg\mathbf{Heard}_i\mathbf{Heard}_i\varphi$.
6. If $\mathrm{M}, s \models^{\mathrm{I}} \mathbf{Saw}_i\varphi$, then, by item 4, also $\mathrm{M}, s \models^{\mathrm{I}} \mathbf{B}_i^o\mathbf{Saw}_i\varphi$. If furthermore $\mathrm{M}, s \models^{\mathrm{I}} \mathbf{M}_i\neg\mathbf{Saw}_i\varphi$ it directly follows that $\mathrm{M}, s \models^{\mathrm{I}} \mathbf{Saw}_i\mathbf{Saw}_i\varphi$, which suffices to conclude that the left-to-right implication holds. The right-to-left implication follows by the definition of $\mathbf{Saw}_i\varphi$ and the fact that observational beliefs are veridical.

Items 7, 8 and 9 are proved in a similar way as 4, 5 and 6, respectively.

10. Assume $\mathrm{M}, s \models^{\mathrm{I}} \mathbf{Heard}_i\varphi \wedge \mathbf{Heard}_i\psi$. Then $\mathrm{M}, s \models^{\mathrm{I}} \mathbf{B}_i^c\varphi \wedge \neg\mathbf{B}_i^o\varphi \wedge \mathbf{B}_i^c\psi \wedge \neg\mathbf{B}_i^o\psi$, and thus $\mathrm{M}, s \models^{\mathrm{I}} \mathbf{B}_i^c(\varphi \wedge \psi) \wedge \neg\mathbf{B}_i^o(\varphi \wedge \psi)$. Hence $\mathrm{M}, s \models^{\mathrm{I}} \mathbf{Heard}_i(\varphi \wedge \psi)$.
11. If ψ is a tautology, i.e. $\models^{\mathrm{I}} \psi$ holds, then also $\models^{\mathrm{I}} \mathbf{K}_i\psi$. It is easy to see that the formula $\mathbf{Heard}_i(\varphi \wedge \psi) \wedge \mathbf{Heard}_i\varphi \wedge \neg\mathbf{Heard}_i\psi$ is satisfiable for some formula φ which is contingent, i.e. a formula φ which is neither a tautology nor a contradiction.
12. Assume $\mathrm{M}, s \models^{\mathrm{I}} \mathbf{Heard}_i(\varphi \wedge \psi)$. Since $\models^{\mathrm{I}} (\varphi \wedge \psi) \to \varphi$ and $\models^{\mathrm{I}} (\varphi \wedge \psi) \to \psi$ both hold, the first of the conjuncts on the right-hand side follows more or less directly from the RM'-rule, which is proved below. For the second conjunct note that $\mathrm{M}, s \models^{\mathrm{I}} \neg\mathbf{Heard}_i\varphi \wedge \neg\mathbf{Heard}_i\psi$ together with $\mathrm{M}, s \models^{\mathrm{I}} \mathbf{Heard}_i(\varphi \wedge \psi)$ would imply $\mathrm{M}, s \models^{\mathrm{I}} \mathbf{B}_i^o(\varphi \wedge \psi)$ which contradicts $\mathrm{M}, s \models^{\mathrm{I}} \mathbf{Heard}_i(\varphi \wedge \psi)$.
13. If $\models^{\mathrm{I}} \varphi$ then also $\models^{\mathrm{I}} \mathbf{K}_i\varphi \wedge \mathbf{B}_i^o\varphi \wedge \mathbf{B}_i^c\varphi \wedge \mathbf{B}_i^d\varphi$, and hence directly $\models^{\mathrm{I}} \neg\mathbf{Saw}_i\varphi \wedge \neg\mathbf{Heard}_i\varphi \wedge \neg\mathbf{Jumped}_i\varphi$.
14. If ψ is a tautology, then so is $\varphi \to \psi$, and, by the previous clause, $\models^{\mathrm{I}} \neg\mathbf{Heard}_i\psi$.

Then for an appropriate contingency φ, the formula $\mathbf{Heard}_i\varphi \wedge \neg\mathbf{Heard}_i\psi$ is easily satisfiable.

15. Assume that $\models^{\mathrm{I}} (\varphi \to \psi)$ and $\mathrm{M}, s \models^{\mathrm{I}} \mathbf{Heard}_i\varphi$, i.e. $\mathrm{M}, s \models^{\mathrm{I}} \mathbf{B}^c_i\varphi \wedge \neg\mathbf{B}^o_i\varphi$. Then also $\mathrm{M}, s \models^{\mathrm{I}} \mathbf{B}^c_i\psi$. Now if $\mathrm{M}, s \not\models^{\mathrm{I}} \mathbf{B}^o_i\psi$, then $\mathrm{M}, s \models^{\mathrm{I}} \mathbf{Heard}_i\psi$; otherwise if $\mathrm{M}, s \not\models^{\mathrm{I}} \mathbf{K}_i\psi$ it holds that $\mathrm{M}, s \models^{\mathrm{I}} \mathbf{Saw}_i\psi$, and else $\mathrm{M}, s \models^{\mathrm{I}} \mathbf{K}_i\psi$. Hence $\mathrm{M}, s \models^{\mathrm{I}} \bigvee_{\mathbf{Z} \succeq' \mathbf{Heard}_i} \mathbf{Z}\psi$.
16. Suppose $\models^{\mathrm{I}} \varphi \leftrightarrow \psi$ and $\mathrm{M}, s \models^{\mathrm{I}} \mathbf{Heard}_i\varphi$. Then $\mathrm{M}, s \models^{\mathrm{I}} \mathbf{B}^c_i\varphi \wedge \neg\mathbf{B}^o_i\varphi$ and thus $\mathrm{M}, s \models^{\mathrm{I}} \mathbf{B}^c_i\psi \wedge \neg\mathbf{B}^o_i\psi$. Hence $\mathrm{M}, s \models^{\mathrm{I}} \mathbf{Heard}_i\psi$.

⊠

5.18. Proposition. *For all $\alpha \in \mathrm{Ac}^{\mathrm{I}}$ and $\varphi \in \mathrm{L}^{\mathrm{I}}$ we have:*

- *if α is realisable and x-informative with regard to φ then α is genuinely x-informative with regard to φ.*
- *if α is deterministic and genuinely x-informative with regard to φ then α is x-informative with regard to φ.*

Proof: We show both items. Let $\alpha \in \mathrm{Ac}^{\mathrm{I}}$ and $\varphi \in \mathrm{L}^{\mathrm{I}}$ be arbitrary.

- If α is realisable and x-informative with regard to φ, then $\mathbf{F}^{\mathrm{I}} \models^{\mathrm{I}} \langle \mathrm{do}_i(\alpha)\rangle\top$ and $\mathbf{F}^{\mathrm{I}} \models^{\mathrm{I}} [\mathrm{do}_i(\alpha)]\mathbf{Bwhether}^x_i\varphi$ as schemas in $i \in \mathrm{A}$. But then also $\mathbf{F}^{\mathrm{I}} \models^{\mathrm{I}} \langle \mathrm{do}_i(\alpha)\rangle\mathbf{Bwhether}^x_i\varphi$ as a schema in $i \in \mathrm{A}$, hence α is genuinely x-informative with regard to φ.
- If α is deterministic and genuinely x-informative with regard to φ, then $\mathbf{F}^{\mathrm{I}}$ satisfies the schema $\langle \mathrm{do}_i(\alpha)\rangle\psi \to [\mathrm{do}_i(\alpha)]\psi$ in $i \in \mathrm{A}$ and $\psi \in \mathrm{L}^{\mathrm{I}}$. Furthermore, $\mathbf{F}^{\mathrm{I}}$ satisfies the schema $\langle \mathrm{do}_i(\alpha)\rangle\mathbf{Bwhether}^x_i\varphi$ in $i \in \mathrm{A}$. But then $\mathbf{F}^{\mathrm{I}}$ satisfies the schema $[\mathrm{do}_i(\alpha)]\mathbf{Bwhether}^x_i\varphi$ in $i \in \mathrm{A}$, i.e. α is x-informative with regard to φ.

⊠

5.20. Proposition. *For all $\mathrm{M} \in \mathbf{M}^{\mathrm{I}}$, $s \in \mathrm{M}$, $i \in \mathrm{A}$, $f \in \mathrm{L}_0$ it holds for $x \in \{c, d\}$:*

- $\mathtt{revise}^o(i, f)(\mathrm{M}, s) \in \mathbf{M}^{\mathrm{I}}$
- *if* $\mathtt{revise}^x(i, f)(\mathrm{M}, s)$ *is defined, then* $\mathtt{revise}^x(i, f)(\mathrm{M}, s) \in \mathbf{M}^{\mathrm{I}}$

Proof: We show the first item, leaving the second item, which is analogous to the first one but considerably simpler, to the reader. Let $\mathrm{M} = \langle \mathrm{S}, \pi, \mathrm{R}, \mathrm{B}^o, \mathrm{B}^c, \mathrm{B}^d, \mathrm{D}, \mathtt{r}_0, \mathtt{c}_0\rangle \in \mathbf{M}^{\mathrm{I}}$ with state s, $i \in \mathrm{A}$ and $f \in \mathrm{L}_0$ be arbitrary. Let M' be the tuple such that $\mathrm{M}' = \mathtt{revise}^o(i, f)(\mathrm{M}, s)$. Then M' is of the form $\langle \mathrm{S}, \pi, \mathrm{R}, \mathrm{B}^{o\prime}, \mathrm{B}^{c\prime}, \mathrm{B}^{d\prime}, \mathrm{D}, \mathtt{r}_0, \mathtt{c}_0\rangle$. To show that $\mathrm{M}' \in \mathbf{M}^{\mathrm{I}}$ we have to show for all $j \in \mathrm{A}$ and states t, t' in M that

1. $\mathrm{B}^{o\prime}(j)$ is an equivalence relation
2. $\mathrm{B}^{d\prime}(j, t) \neq \emptyset$
3. $\mathrm{B}^{d\prime}(j, t) \subseteq \mathrm{B}^{c\prime}(j, t) \subseteq [t]_{\mathrm{B}^{o\prime}(j)} \subseteq [t]_{\mathrm{R}(j)}$
4. if $t' \in [t]_{\mathrm{B}^{o\prime}(j)}$ then $\mathrm{B}^{c\prime}(j, t') = \mathrm{B}^{c\prime}(j, t)$ and $\mathrm{B}^{d\prime}(j, t') = \mathrm{B}^{d\prime}(j, t)$

We successively show that M' indeed meets these four requirements.

1. If $j \neq i$ then $\mathrm{B}^{o'}(j) = \mathrm{B}^{o}(j)$, and hence $\mathrm{B}^{o'}(j)$ is indeed an equivalence relation, so it suffices to show that $\mathrm{B}^{o'}(i)$ is an equivalence relation. From Definition 5.19 it follows that $\mathrm{B}^{o'}(i) = (\mathrm{B}^{o}(i) \setminus \mathrm{Cl}_{\mathrm{eq}}([s]_{\mathrm{B}^{o}(i)})) \cup \mathrm{Cl}_{\mathrm{eq}}([s]^{f}_{\mathrm{B}^{o}(i)}) \cup \mathrm{Cl}_{\mathrm{eq}}([s]^{\neg f}_{\mathrm{B}^{o}(i)})$. We show that $\mathrm{B}^{o'}(i)$ is an equivalence relation by showing that it is reflexive, symmetrical and transitive.
   - It is obvious that $(s', s') \in \mathrm{B}^{o'}(i)$ for all $s' \in \mathrm{S}$. For if $s' \notin [s]_{\mathrm{B}^{o}(i)}$, $(s', s') \in \mathrm{B}^{o}(i) \setminus \mathrm{Cl}_{\mathrm{eq}}([s]_{\mathrm{B}^{o}(i)})$, and otherwise s' is either in $[s]^{f}_{\mathrm{B}^{o}(i)}$ or in $[s]^{\neg f}_{\mathrm{B}^{o}(i)}$, which implies that either $(s', s') \in \mathrm{Cl}_{\mathrm{eq}}([s]^{f}_{\mathrm{B}^{o}(i)})$ or $(s', s') \in \mathrm{Cl}_{\mathrm{eq}}([s]^{\neg f}_{\mathrm{B}^{o}(i)})$. In either case $(s', s') \in \mathrm{B}^{o'}(i)$ and hence $\mathrm{B}^{o'}(i)$ is reflexive.
   - Let (s_1, s_2) in $\mathrm{B}^{o'}(i)$. If (s_1, s_2) is either in $\mathrm{Cl}_{\mathrm{eq}}([s]^{f}_{\mathrm{B}^{o}(i)})$ or in $\mathrm{Cl}_{\mathrm{eq}}([s]^{\neg f}_{\mathrm{B}^{o}(i)})$ it follows by definition of $\mathrm{Cl}_{\mathrm{eq}}$ that either $(s_2, s_1) \in \mathrm{Cl}_{\mathrm{eq}}([s]^{f}_{\mathrm{B}^{o}(i)})$ or $(s_2, s_1) \in \mathrm{Cl}_{\mathrm{eq}}([s]^{\neg f}_{\mathrm{B}^{o}(i)})$, hence $(s_2, s_1) \in \mathrm{B}^{o'}(i)$. So assume that $(s_1, s_2) \in \mathrm{B}^{o}(i) \setminus \mathrm{Cl}_{\mathrm{eq}}([s]_{\mathrm{B}^{o}(i)})$. Then $(s_1, s_2) \in \mathrm{B}^{o}(i)$ and $(s_1, s_2) \notin \mathrm{Cl}_{\mathrm{eq}}([s]_{\mathrm{B}^{o}(i)})$. Since $\mathrm{B}^{o}(i)$ is symmetrical, it follows that $(s_2, s_1) \in \mathrm{B}^{o}(i)$, and, by definition of $\mathrm{Cl}_{\mathrm{eq}}$, we have that $(s_2, s_1) \notin \mathrm{Cl}_{\mathrm{eq}}([s]_{\mathrm{B}^{o}(i)})$. Hence $(s_2, s_1) \in \mathrm{B}^{o}(i) \setminus \mathrm{Cl}_{\mathrm{eq}}([s]_{\mathrm{B}^{o}(i)}) \subseteq \mathrm{B}^{o'}(i)$ which suffices to conclude that $\mathrm{B}^{o'}(i)$ is symmetrical.
   - Let $(s_1, s_2) \in \mathrm{B}^{o'}(i)$ and $(s_2, s_3) \in \mathrm{B}^{o'}(i)$. We distinguish three cases:
     - $(s_1, s_2) \in \mathrm{Cl}_{\mathrm{eq}}([s]^{f}_{\mathrm{B}^{o}(i)})$. Then $s_2 \in [s]^{f}_{\mathrm{B}^{o}(i)}$. From $(s_2, s_3) \in \mathrm{B}^{o'}(i)$ it follows that also $s_3 \in [s]^{f}_{\mathrm{B}^{o}(i)}$. Hence $(s_2, s_3) \in \mathrm{Cl}_{\mathrm{eq}}([s]^{f}_{\mathrm{B}^{o}(i)})$. By definition of $\mathrm{Cl}_{\mathrm{eq}}$ it follows that $(s_1, s_3) \in \mathrm{Cl}_{\mathrm{eq}}([s]^{f}_{\mathrm{B}^{o}(i)}) \subseteq \mathrm{B}^{o'}(i)$.
     - $(s_1, s_2) \in \mathrm{Cl}_{\mathrm{eq}}([s]^{\neg f}_{\mathrm{B}^{o}(i)})$. This case is completely analogous to the case where $(s_1, s_2) \in \mathrm{Cl}_{\mathrm{eq}}([s]^{f}_{\mathrm{B}^{o}(i)})$.
     - $(s_1, s_2) \in \mathrm{B}^{o}(i) \setminus \mathrm{Cl}_{\mathrm{eq}}([s]_{\mathrm{B}^{o}(i)})$. In this case $(s_1, s_2) \in \mathrm{B}^{o}(i)$ and $(s_1, s_2) \notin \mathrm{Cl}_{\mathrm{eq}}([s]_{\mathrm{B}^{o}(i)})$. Since $\mathrm{B}^{o}(i)$ is an equivalence relation this implies that $s_1 \notin [s]_{\mathrm{B}^{o}(i)}$ and $s_2 \notin [s]_{\mathrm{B}^{o}(i)}$: for if either one of them would be in $[s]_{\mathrm{B}^{o}(i)}$, they would both be, due to the transitivity of $\mathrm{B}^{o}(i)$, and this contradicts $(s_1, s_2) \notin \mathrm{Cl}_{\mathrm{eq}}([s]_{\mathrm{B}^{o}(i)})$. From $s_2 \notin [s]_{\mathrm{B}^{o}(i)}$ it follows that $(s_2, s_3) \notin \mathrm{Cl}_{\mathrm{eq}}([s]_{\mathrm{B}^{o}(i)})$, and thus $(s_2, s_3) \in \mathrm{B}^{o}(i) \setminus \mathrm{Cl}_{\mathrm{eq}}([s]_{\mathrm{B}^{o}(i)})$. But then $(s_1, s_3) \in \mathrm{B}^{o}(i)$ and since $s_1 \notin [s]_{\mathrm{B}^{o}(i)}$ it follows that $(s_1, s_3) \in \mathrm{B}^{o}(i) \setminus \mathrm{Cl}_{\mathrm{eq}}([s]_{\mathrm{B}^{o}(i)}) \subseteq \mathrm{B}^{o'}(i)$.

     Hence $\mathrm{B}^{o'}(i)$ is transitive.

   Thus $\mathrm{B}^{o'}(i)$ is reflexive, symmetrical and transitive, and hence an equivalence relation.
2. If $j \neq i$ or $t \notin [s]_{\mathrm{B}^{o}(i)}$ then $\mathrm{B}^{d'}(j, t) = \mathrm{B}^{d}(j, t)$, and since $\mathrm{M} \in \mathbf{M}^{\mathrm{I}}$, $\mathrm{B}^{d}(j, t) \neq \emptyset$. So let $t \in [s]_{\mathrm{B}^{o}(i)}$. By definition either $\mathrm{B}^{d'}(i, t) = \mathrm{B}^{d}(i, t) \cap [t]_{\mathrm{B}^{o'}(i)}$, in which case the latter set is demanded to be non-empty, or $\mathrm{B}^{d'}(i, t) = \mathrm{B}^{c'}(i, t)$. Now either $\mathrm{B}^{c'}(i, t) = \mathrm{B}^{c}(i, t) \cap [t]_{\mathrm{B}^{o'}(i)}$, in which case the latter set is demanded to be non-empty, or $\mathrm{B}^{c'}(i, t) = [t]_{\mathrm{B}^{o'}(i)}$, which, since $\mathrm{B}^{o'}(i)$ is an equivalence relation, is also not empty. Hence also if $\mathrm{B}^{d'}(i, t) = \mathrm{B}^{c'}(i, t)$ it is ensured to be non-empty. Thus $\mathrm{B}^{d}(i, t) \neq \emptyset$ for all $t \in [s]_{\mathrm{B}^{o}(i)}$.

3. If $j \neq i$ or $t \notin [s]_{\mathrm{B}^{o}(i)}$ then this requirement is trivially met, since in this case M' is no different than M. So consider the case where $t \in [s]_{\mathrm{B}^{o}(i)}$. If $\mathrm{B}^{d'}(i,t) = \mathrm{B}^{c'}(i,t)$ then directly $\mathrm{B}^{d'}(i,t) \subseteq \mathrm{B}^{c'}(i,t)$. So let $\mathrm{B}^{d'}(i,t) = \mathrm{B}^{d}(i,t) \cap [t]_{\mathrm{B}^{o'}(i)}$. Now if $\mathrm{B}^{c'}(i,t) = \mathrm{B}^{c}(i,t) \cap [t]_{\mathrm{B}^{o'}(i)}$, then, since $\mathrm{B}^{d}(i,t) \subseteq \mathrm{B}^{c}(i,t)$, we have that $\mathrm{B}^{d'}(i,t) \subseteq \mathrm{B}^{c'}(i,t)$. Else $\mathrm{B}^{c'}(i,t) = [t]_{\mathrm{B}^{o'}(i)}$ which, since $\mathrm{B}^{d'}(i,t) = \mathrm{B}^{d}(i,t) \cap [t]_{\mathrm{B}^{o'}(i)}$, also implies that $\mathrm{B}^{d'}(i,t) \subseteq \mathrm{B}^{c'}(i,t)$. Thus $\mathrm{B}^{d'}(i,t) \subseteq \mathrm{B}^{c'}(i,t)$. From the definition of $\mathrm{B}^{c'}(i,t)$ it follows directly that $\mathrm{B}^{c'}(i,t) \subseteq [t]_{\mathrm{B}^{o'}(i)}$, which leaves only to show that $[t]_{\mathrm{B}^{o'}(i)} \subseteq [t]_{\mathrm{R}(i)}$. Since $t \in [s]_{\mathrm{B}^{o}(i)}$ we have by Definition 5.19 that either $[t]_{\mathrm{B}^{o'}(i)} = [s]^{f}_{\mathrm{B}^{o}(i)}$ or $[t]_{\mathrm{B}^{o'}(i)} = [s]^{\neg f}_{\mathrm{B}^{o}(i)}$. Since $t \in [s]_{\mathrm{B}^{o}(i)}$ it follows that $[t]_{\mathrm{B}^{o}(i)} = [s]_{\mathrm{B}^{o}(i)}$, and since both $[s]^{f}_{\mathrm{B}^{o}(i)} \subseteq [s]_{\mathrm{B}^{o}(i)}$ and $[s]^{\neg f}_{\mathrm{B}^{o}(i)} \subseteq [s]_{\mathrm{B}^{o}(i)}$ it follows that $[t]_{\mathrm{B}^{o'}(i)} \subseteq [t]_{\mathrm{B}^{o}(i)}$. Since $\mathrm{M} \in \mathbf{M}^{\mathrm{I}}$ we know that $[t]_{\mathrm{B}^{o}(i)} \subseteq [t]_{\mathrm{R}(i)}$, which suffices to conclude that $[t]_{\mathrm{B}^{o'}(i)} \subseteq [t]_{\mathrm{R}(i)}$.
4. If either $j \neq i$ or $t \notin [s]_{\mathrm{B}^{o}(i)}$ then this requirement is trivially met. So let $t \in [s]_{\mathrm{B}^{o}(i)}$ and assume that $t' \in [t]_{\mathrm{B}^{o'}(i)}$. Note that this implies that $[t']_{\mathrm{B}^{o'}(i)} = [t]_{\mathrm{B}^{o'}(i)}$. Furthermore, if $t' \in [t]_{\mathrm{B}^{o'}(i)}$ then either $\{t,t'\} \subseteq [s]^{f}_{\mathrm{B}^{o}(i)}$ or $\{t,t'\} \subseteq [s]^{\neg f}_{\mathrm{B}^{o}(i)}$, which in particular implies that $\{t,t'\} \subseteq [s]_{\mathrm{B}^{o}(i)}$. Since $\mathrm{M} \in \mathbf{M}^{\mathrm{I}}$ we have that $\mathrm{B}^{c}(i,t') = \mathrm{B}^{c}(i,t)$ and $\mathrm{B}^{d}(i,t') = \mathrm{B}^{d}(i,t)$. This implies that $\mathrm{B}^{c}(i,t') \cap [t']_{\mathrm{B}^{o'}(i)} = \mathrm{B}^{c}(i,t) \cap [t]_{\mathrm{B}^{o'}(i)}$ and $[t']_{\mathrm{B}^{o'}(i)} = [t]_{\mathrm{B}^{o'}(i)}$, which suffices to conclude that $\mathrm{B}^{c'}(i,t') = \mathrm{B}^{c'}(i,t)$. In combination with $\mathrm{B}^{d}(i,t') \cap [t']_{\mathrm{B}^{o'}(i)} = \mathrm{B}^{d}(i,t) \cap [t]_{\mathrm{B}^{o'}(i)}$ this yields that $\mathrm{B}^{d'}(i,t') = \mathrm{B}^{d'}(i,t)$.

Since M' meets all four requirements we conclude that it is indeed an element of $\mathbf{M}^{\mathrm{I}}$.
⊠

5.23. Proposition. *For all* $\mathrm{M} \in \mathbf{M}^{\mathrm{I}}$, $s \in \mathrm{M}$, $i \in \mathrm{A}$, $f, g \in \mathrm{L}_0$ *and* $x \in \{c,d\}$ *we have:*

1. $f \in \mathrm{B}^{\star f}_{x}(i,\mathrm{M},s)$
2. $\mathrm{B}^{\star f}_{x}(i,\mathrm{M},s) \subseteq \mathrm{B}^{+f}_{x}(i,\mathrm{M},s)$
3. *If* $\neg f \notin \mathrm{B}_{x}(i,\mathrm{M},s)$ *then* $\mathrm{B}^{+f}_{x}(i,\mathrm{M},s) \subseteq \mathrm{B}^{\star f}_{x}(i,\mathrm{M},s)$
4. $\mathrm{B}^{\star f}_{x}(i,\mathrm{M},s) = \mathrm{B}^{\perp}$ *if and only if* $\mathrm{M}, s \models^{\mathrm{I}} \mathbf{B}^{suc(x)}_{i} \neg f$
5. *If* $\mathrm{M}, s \models^{\mathrm{I}} \mathbf{B}^{suc(x)}_{i}(f \leftrightarrow g)$ *then* $\mathrm{B}^{\star f}_{x}(i,\mathrm{M},s) = \mathrm{B}^{\star g}_{x}(i,\mathrm{M},s)$
6. $\mathrm{B}^{\star f \wedge g}_{x}(i,\mathrm{M},s) \subseteq \mathrm{B}^{\star f+g}_{x}(i,\mathrm{M},s)$
7. *If* $\neg g \notin \mathrm{B}^{\star f}_{x}(i,\mathrm{M},s)$ *then* $\mathrm{B}^{\star f+g}_{x}(i,\mathrm{M},s) \subseteq \mathrm{B}^{\star f \wedge g}_{x}(i,\mathrm{M},s)$

Proof: Let $\mathrm{M} \in \mathbf{M}^{\mathrm{I}}$ with state s, $i \in \mathrm{A}$ and $f, g \in \mathrm{L}_0$ be arbitrary. We successively show all items of Proposition 5.23 for the case that $x = c$; the case where $x = d$ is completely analogous.

1. If $\mathtt{revise}^{c}(i,f)(\mathrm{M},s) \in \mathbf{M}^{\mathrm{I}}$, then either $\mathrm{B}^{c'}(i,s) = \mathrm{B}^{c}(i,s) \cap [\![f]\!]$ or $\mathrm{B}^{c'}(i,s) = [s]_{\mathrm{B}^{o}(i)} \cap [\![f]\!]$. In both cases, $\mathrm{B}^{c'}(i,s) \subseteq [\![f]\!]$, and thus $\mathrm{M}', s \models^{\mathrm{I}} \mathbf{B}^{c}_{i} f$, i.e. $f \in \mathrm{B}^{\star f}_{c}(i,\mathrm{M},s)$. If $\mathtt{revise}^{c}(i,f)(\mathrm{M},s)$ is undefined, then $\mathrm{B}^{\star f}_{c}(i,\mathrm{M},s) = \mathrm{L}_0$ and thus $f \in \mathrm{B}^{\star f}_{c}(i,\mathrm{M},s)$.
2. If $\mathrm{B}^{c}(i,s) \cap [\![f]\!] = \emptyset$ then $\neg f \in \mathrm{B}_{c}(i,\mathrm{M},s)$, hence $\mathrm{B}^{+f}_{c}(i,\mathrm{M},s) = \mathrm{L}_0$, and thus $\mathrm{B}^{\star f}_{c}(i,\mathrm{M},s) \subseteq \mathrm{B}^{+f}_{c}(i,\mathrm{M},s)$. Hence assume that $\mathrm{B}^{c}(i,s) \cap [\![f]\!] \neq \emptyset$. Now $\mathrm{B}^{\star f}_{c}(i,\mathrm{M},s) = \{g \in \mathrm{L}_0 \mid \mathrm{M}', s' \models^{\mathrm{I}} g \text{ for all } s' \in \mathrm{B}^{c'}(i,s)\}$. Since $\mathrm{B}^{c}(i,s) \cap [\![f]\!] \neq \emptyset$, $\mathrm{B}^{c'}(i,s) =$

$\mathrm{B}^c(i,s) \cap [\![f]\!]$. Now since f and g are purely propositional, we have that $\mathrm{M}', s' \models^{\mathrm{I}} g$ for all $s' \in \mathrm{B}^{c'}(i,s)$ iff $\mathrm{M}, s' \models^{\mathrm{I}} (f \to g)$ for all $s' \in \mathrm{B}^c(i,s)$. Thus $g \in \mathrm{B}_c^{\star f}(i,\mathrm{M},s)$ iff $(f \to g) \in \mathrm{B}_c(i,\mathrm{M},s)$. But then $g \in \mathrm{B}_c^{\star f}(i,\mathrm{M},s)$ implies $g \in \mathrm{Cn}(\mathrm{B}^c(i,\mathrm{M},s) \cup \{f\})$, which suffices to conclude that $\mathrm{B}_c^{\star f}(i,\mathrm{M},s) \subseteq \mathrm{B}_c^{+f}(i,\mathrm{M},s)$.

3. As shown in the previous item, if $\mathrm{B}^c(i,s) \cap [\![f]\!] \neq \emptyset$, then $\mathrm{B}_c^{\star f}(i,\mathrm{M},s) = \{g \in \mathrm{L}_0 \mid \mathrm{M}, s \models^{\mathrm{I}} \mathbf{B}_i^c(f \to g)\}$. Hence we have to show that $\mathrm{Cn}(\mathrm{B}_c(i,\mathrm{M},s) \cup \{f\})$ is contained in $\{g \in \mathrm{L}_0 \mid \mathrm{M}, s \models^{\mathrm{I}} \mathbf{B}_i^c(f \to g)\}$. By the deduction theorem for classical propositional logic we have that $g \in \mathrm{Cn}(\mathrm{B}_c(i,\mathrm{M},s) \cup \{f\})$ iff $(f \to g) \in \mathrm{Cn}(\mathrm{B}_c(i,\mathrm{M},s))$. Since $\mathrm{B}_c(i,\mathrm{M},s)$ is closed under the deduction of classical propositional logic, $\mathrm{Cn}(\mathrm{B}_c(i,\mathrm{M},s)) = \mathrm{B}_c(i,\mathrm{M},s)$, which suffices to conclude item 3.
4. From Definition 5.19 it follows that $\mathtt{revise}^c(i,f)(\mathrm{M},s)$ is undefined iff $\mathrm{M}, s \models^{\mathrm{I}} \mathbf{B}_i^o \neg f$. Hence $\mathrm{B}_c^{\star f}(i,\mathrm{M},s) = \mathrm{B}^\bot$ iff $\mathrm{M}, s \models^{\mathrm{I}} \mathbf{B}_i^o \neg f$, i.e. iff $\mathrm{M}, s \models^{\mathrm{I}} \mathbf{B}_i^{suc(c)} \neg f$.
5. If $\mathrm{M}, s \models^{\mathrm{I}} \mathbf{B}_i^o(f \leftrightarrow g)$ then both $\mathrm{B}^c(i,s) \cap [\![f]\!] = \mathrm{B}^c(i,s) \cap [\![g]\!]$ and $[s]_{\mathrm{B}^o(i)} \cap [\![f]\!] = [s]_{\mathrm{B}^o(i)} \cap [\![g]\!]$, which suffices to conclude that item 5 holds.
6. To show this item we distinguish six cases:
   1. $[s]_{\mathrm{B}^o(i)} \cap [\![f]\!] = \emptyset$. In this case $\mathtt{revise}^c(i,f)(\mathrm{M},s)$ is undefined. Hence $\mathrm{B}_c^{\star f}(i,\mathrm{M},s) = \mathrm{B}^\bot$ and thus $\mathrm{B}_c^{\star f+g}(i,\mathrm{M},s) = \mathrm{B}^\bot$, which suffices to conclude that $\mathrm{B}_c^{\star f\wedge g}(i,\mathrm{M},s) \subseteq \mathrm{B}_c^{\star f+g}(i,\mathrm{M},s)$.
   2. $\mathrm{B}^c(i,s) \cap [\![f]\!] = \emptyset, [s]_{\mathrm{B}^o(i)} \cap [\![f]\!] \neq \emptyset, [s]_{\mathrm{B}^o(i)} \cap [\![f \wedge g]\!] = \emptyset$. In this case $\mathrm{B}_c^{\star f}(i,\mathrm{M},s) = \{h \in \mathrm{L}_0 \mid \mathrm{M}, s \models^{\mathrm{I}} \mathbf{B}_i^o(f \to h)\}$. Since $[s]_{\mathrm{B}^o(i)} \cap [\![f \wedge g]\!] = \emptyset$, $\mathrm{M}, s \models^{\mathrm{I}} \mathbf{B}_i^o(f \to \neg g)$. Thus $\neg g \in \mathrm{B}_c^{\star f}(i,\mathrm{M},s)$, and hence $\mathrm{B}_c^{\star f+g}(i,\mathrm{M},s) = \mathrm{B}^\bot$. Thus $\mathrm{B}_c^{\star f\wedge g}(i,\mathrm{M},s) \subseteq \mathrm{B}_c^{\star f+g}(i,\mathrm{M},s)$.
   3. $\mathrm{B}^c(i,s) \cap [\![f]\!] \neq \emptyset, [s]_{\mathrm{B}^o(i)} \cap [\![f \wedge g]\!] = \emptyset$. In this case $\mathrm{B}_c^{\star f}(i,\mathrm{M},s) = \{h \in \mathrm{L}_0 \mid \mathrm{M}, s \models^{\mathrm{I}} \mathbf{B}_i^c(f \to h)\}$. Since $[s]_{\mathrm{B}^o(i)} \cap [\![f \wedge g]\!] = \emptyset$, also $\mathrm{B}^c(i,s) \cap [\![f \wedge g]\!] = \emptyset$. Thus $\mathrm{M}, s \models^{\mathrm{I}} \mathbf{B}_i^c(f \to \neg g)$, and therefore $\neg g \in \mathrm{B}_c^{\star f}(i,\mathrm{M},s)$. Then $\mathrm{B}_c^{\star f+g}(i,\mathrm{M},s) = \mathrm{B}^\bot$, which suffices to conclude that $\mathrm{B}_c^{\star f\wedge g}(i,\mathrm{M},s) \subseteq \mathrm{B}_c^{\star f+g}(i,\mathrm{M},s)$.
   4. $\mathrm{B}^c(i,s) \cap [\![f]\!] = \emptyset, [s]_{\mathrm{B}^o(i)} \cap [\![f \wedge g]\!] \neq \emptyset$. In this case both $\mathtt{revise}^c(i,f)(\mathrm{M}) \in \mathbf{M}^{\mathrm{I}}$ and $\mathtt{revise}^c(i,f \wedge g)(\mathrm{M},s) \in \mathbf{M}^{\mathrm{I}}$. Then $\mathrm{B}_c^{\star f\wedge g}(i,\mathrm{M},s) = \{h \in \mathrm{L}_0 \mid \mathrm{M}, s \models^{\mathrm{I}} \mathbf{B}_i^o((f \wedge g) \to h)\}$ and $\mathrm{B}_c^{\star f}(i,\mathrm{M},s) = \{h \in \mathrm{L}_0 \mid \mathrm{M}, s \models^{\mathrm{I}} \mathbf{B}_i^o(f \to h)\}$. Now let $h \in \mathrm{L}_0$ be arbitrary such that $h \in \mathrm{B}_c^{\star f\wedge g}(i,\mathrm{M},s)$, i.e. $\mathrm{M}, s \models^{\mathrm{I}} \mathbf{B}_i^o((f \wedge g) \to h)$. Then also $\mathrm{M}, s \models^{\mathrm{I}} \mathbf{B}_i^o(f \to (g \to h))$, and hence $(g \to h) \in \mathrm{B}_c^{\star f}(i,\mathrm{M},s)$. Then $h \in \mathrm{Cn}(\mathrm{B}_c^{\star f}(i,\mathrm{M},s) \cup \{g\})$, and thus $h \in \mathrm{B}_c^{\star f+g}(i,\mathrm{M},s)$. Since h is arbitrary it follows that $\mathrm{B}_c^{\star f\wedge g}(i,\mathrm{M},s) \subseteq \mathrm{B}_c^{\star f+g}(i,\mathrm{M},s)$.
   5. $\mathrm{B}^c(i,s) \cap [\![f]\!] \neq \emptyset, \mathrm{B}^c(i,s) \cap [\![f \wedge g]\!] = \emptyset, [s]_{\mathrm{B}^o(i)} \cap [\![f \wedge g]\!] \neq \emptyset$. In this case $\mathrm{B}_c^{\star f\wedge g}(i,\mathrm{M},s)$ is again equal to $\{h \in \mathrm{L}_0 \mid \mathrm{M}, s \models^{\mathrm{I}} \mathbf{B}_i^o((f \wedge g) \to h)\}$ while $\mathrm{B}_c^{\star f}(i,\mathrm{M},s) = \{h \in \mathrm{L}_0 \mid \mathrm{M}, s \models^{\mathrm{I}} \mathbf{B}_i^c(f \to h)\}$. Since $\neg g \in \mathrm{B}_c^{\star f}(i,\mathrm{M},s)$, which follows from $\mathrm{B}^c(i,s) \cap [\![f \wedge g]\!] = \emptyset$, $\mathrm{B}_c^{\star f+g}(i,\mathrm{M},s) = \mathrm{B}^\bot$, and thus $\mathrm{B}_c^{\star f\wedge g}(i,\mathrm{M},s) \subseteq \mathrm{B}_c^{\star f+g}(i,\mathrm{M},s)$.
   6. $\mathrm{B}^c(i,s) \cap [\![f \wedge g]\!] \neq \emptyset$. Then $\mathrm{B}_c^{\star f\wedge g}(i,\mathrm{M},s) = \{h \in \mathrm{L}_0 \mid \mathrm{M}, s \models^{\mathrm{I}} \mathbf{B}_i^c((f \wedge g) \to h)\}$

and $\mathrm{B}_c^{\star f}(i,\mathrm{M},s) = \{h \in \mathrm{L}_0 \mid \mathrm{M},s \models^{\mathrm{I}} \mathbf{B}_i^c(f \to h)\}$. Now if $h \in \mathrm{B}_c^{\star f\wedge g}(i,\mathrm{M},s)$ then $\mathrm{M},s \models^{\mathrm{I}} \mathbf{B}_i^c((f \wedge g) \to h)$, thus $\mathrm{M},s \models^{\mathrm{I}} \mathbf{B}_i^c(f \to (g \to h))$ and hence $(g \to h) \in \mathrm{B}_c^{\star f}(i,\mathrm{M},s)$. Then $h \in \mathrm{Cn}(\mathrm{B}_c^{\star f}(i,\mathrm{M},s) \cup \{g\})$, which suffices to conclude that $\mathrm{B}_c^{\star f\wedge g}(i,\mathrm{M},s) \subseteq \mathrm{B}_c^{\star f+g}(i,\mathrm{M},s)$.

Since these six cases capture all possibilities, we conclude that item 6 of Proposition 5.23 indeed holds.

7. To prove item 7 we distinguish the same six cases as in the previous item. The only two cases that are interesting, are the fourth and the sixth one; for the other cases it holds that $\neg g \in \mathrm{B}_c^{\star f}(i,\mathrm{M},s)$ which trivially proves item 7. Since the interesting cases are completely analogous, we show only the fourth one. So suppose that $\mathrm{B}^c(i,s) \cap [\![f]\!] = \emptyset$ and $[s]_{\mathrm{B}^o(i)} \cap [\![f \wedge g]\!] \neq \emptyset$. As in the previous item we find that $\mathrm{B}_c^{\star f\wedge g}(i,\mathrm{M},s) = \{h \in \mathrm{L}_0 \mid \mathrm{M},s \models^{\mathrm{I}} \mathbf{B}_i^o((f \wedge g) \to h)\}$ and $\mathrm{B}_c^{\star f}(i,\mathrm{M},s) = \{h \in \mathrm{L}_0 \mid \mathrm{M},s \models^{\mathrm{I}} \mathbf{B}_i^o(f \to h)\}$. Now suppose that $h \in \mathrm{B}_c^{\star f+g}(i,\mathrm{M},s)$, i.e. $h \in \mathrm{Cn}(\mathrm{B}_c^{\star f}(i,\mathrm{M},s) \cup \{g\})$. Applying the deduction theorem for classical propositional logic in the same way as in item 3, we find that $(g \to h) \in \mathrm{B}_c^{\star f}(i,\mathrm{M},s)$. Then $\mathrm{M},s \models^{\mathrm{I}} \mathbf{B}_i^o(f \to (g \to h))$, and thus $\mathrm{M},s \models^{\mathrm{I}} \mathbf{B}_i^o((f \wedge g) \to h)$. Therefore $h \in \mathrm{B}_c^{\star f\wedge g}(i,\mathrm{M},s)$, which implies that $\mathrm{B}_c^{\star f+g}(i,\mathrm{M},s) \subseteq \mathrm{B}_c^{\star f\wedge g}(i,\mathrm{M},s)$.

⊠

5.24. Proposition. *For all* $\mathrm{M} \in \mathbf{M}^{\mathrm{I}}$, $s \in \mathrm{M}$, $i \in \mathrm{A}$ *and* $f \in \mathrm{L}_0$ *we have:*

- $\mathrm{M},s \models^{\mathrm{I}} f \Rightarrow \mathrm{B}_o^{\star f}(i,\mathrm{M},s) = \mathrm{B}_o^{+f}(i,\mathrm{M},s)$
- $\mathrm{M},s \models^{\mathrm{I}} \neg f \Rightarrow \mathrm{B}_o^{\star f}(i,\mathrm{M},s) = \mathrm{B}_o^{+\neg f}(i,\mathrm{M},s)$

Proof: Since both cases are completely analogous, we restrict ourselves to proving the first one. So let $\mathrm{M} \in \mathbf{M}^{\mathrm{I}}$ with state s, $i \in \mathrm{A}$ and $f \in \mathrm{L}_0$ be arbitrary. Assume that $\mathrm{M},s \models^{\mathrm{I}} f$ and let $\mathrm{M}' = \mathtt{revise}^o(i,f)(\mathrm{M},s)$. By definition we have that $[s]_{\mathrm{B}^{o'}(i)} = [s]^f_{\mathrm{B}^o(i)} = [s]_{\mathrm{B}^o(i)} \cap [\![f]\!]$. We show that $\mathrm{B}_o^{\star f}(i,\mathrm{M},s) = \mathrm{B}_o^{+f}(i,\mathrm{M},s)$, i.e. $\{g \in \mathrm{L}_0 \mid \mathrm{M}',s \models^{\mathrm{I}} \mathbf{B}_i^o g\} = \mathrm{Cn}(\{g \in \mathrm{L}_0 \mid \mathrm{M},s \models^{\mathrm{I}} \mathbf{B}_i^o g\} \cup \{f\})$, by proving that both sets are contained in one another.

'$\supseteq$' Since $[s]_{\mathrm{B}^{o'}(i)} \subseteq [s]_{\mathrm{B}^o(i)}$, the set of all propositional formulae true at all states from $[s]_{\mathrm{B}^o(i)}$ is contained in the set of all propositional formula true at all states from $[s]_{\mathrm{B}^{o'}(i)}$. Hence $\mathrm{B}_o(i,\mathrm{M},s) \subseteq \mathrm{B}_o^{\star f}(i,\mathrm{M},s)$. Furthermore, since $[s]_{\mathrm{B}^{o'}(i)} \subseteq [\![f]\!]$ we have that f is true at all states from $[s]_{\mathrm{B}^{o'}(i)}$. Hence $f \in \mathrm{B}_o^{\star f}(i,\mathrm{M},s)$. Since $\mathrm{B}_o^{\star f}(i,\mathrm{M},s) = \mathrm{B}_o(i,\mathrm{M}',s)$ is deductively closed, we have $\mathrm{Cn}(\mathrm{B}^o(i,\mathrm{M},s) \cup \{f\}) \subseteq \mathrm{B}_o^{\star f}(i,\mathrm{M},s)$.

'$\subseteq$' Suppose that $h \in \mathrm{B}_o^{\star f}(i,\mathrm{M},s)$, i.e. $\mathrm{M}',s \models^{\mathrm{I}} \mathbf{B}_i^o h$. Since $[s]_{\mathrm{B}^{o'}(i)} = [s]_{\mathrm{B}^o(i)} \cap [\![f]\!]$ it follows that $\mathrm{M},t \models^{\mathrm{I}} h$ for all $t \in [s]_{\mathrm{B}^o(i)} \cap [\![f]\!]$. But this implies that $\mathrm{M},t \models^{\mathrm{I}} (f \to h)$ for all $t \in [s]_{\mathrm{B}^o(i)}$, and thus $\mathrm{M},s \models^{\mathrm{I}} \mathbf{B}_i^o(f \to h)$. Then $(f \to h) \in \mathrm{B}_o(i,\mathrm{M},s)$ and $h \in \mathrm{Cn}(\mathrm{B}_o(i,\mathrm{M},s) \cup \{f\})$. Since h is arbitrary it follows that $\mathrm{B}_o^{\star f}(i,\mathrm{M},s) \subseteq \mathrm{Cn}(\mathrm{B}^o(i,\mathrm{M},s) \cup \{f\})$, which was to be shown.

⊠

5.26. Proposition. *For all* $\mathrm{M} \in \mathbf{M}^{\mathrm{I}}$ *with state* s*, for all* $i \in \mathrm{A}$ *and* $f \in \mathrm{L}_0$*, it holds that if* $\mathrm{M}', s = \mathtt{r}^{\mathrm{I}}(i, \mathtt{observe}\, f)(\mathrm{M}, s)$*, then* $\mathrm{M}' \in \mathrm{M}^{\mathrm{I}}_{\sim}$.

Proof: In the case that $\mathrm{M}, s \models^{\mathrm{I}} \mathbf{Bwhether}^o_i f$, $\mathrm{M}' = \mathrm{M}$ and hence $\mathrm{M}' \in \mathbf{M}^{\mathrm{I}}$. Otherwise, $\mathrm{M}' = \mathtt{revise}^o(i, f)(\mathrm{M}, s)$ and we have, by Proposition 5.20, that $\mathrm{M}' \in \mathbf{M}^{\mathrm{I}}$. Now trivially, $\mathrm{M} \in \mathrm{M}^{\mathrm{I}}_{\sim}$, and inspection of Definition 5.19 shows that also $\mathtt{revise}^o(i, f)(\mathrm{M}, s)$ differs from M only in the B-functions, which implies that $\mathtt{revise}^o(i, f)(\mathrm{M}, s) \in \mathrm{M}^{\mathrm{I}}_{\sim}$.
⊠

5.27. Proposition. *For all* $i, j \in \mathrm{A}$*,* $f, g \in \mathrm{L}_0$ *and* $\varphi \in \mathrm{L}^{\mathrm{I}}$ *we have:*

1. $\mathtt{observe}\, f$ *is* o*-informative and truthful with respect to* f
2. $\mathtt{observe}\, f$ *is deterministic, idempotent and realisable*
3. $\models^{\mathrm{I}} \mathbf{K}_j g \leftrightarrow \langle \mathrm{do}_i(\mathtt{observe}\, f)\rangle \mathbf{K}_j g$
4. $\models^{\mathrm{I}} \mathbf{B}^o_j g \rightarrow \langle \mathrm{do}_i(\mathtt{observe}\, f)\rangle \mathbf{B}^o_j g$
5. $\models^{\mathrm{I}} (f \wedge \langle \mathrm{do}_i(\mathtt{observe}\, f)\rangle \mathbf{B}^o_i g) \leftrightarrow (f \wedge \mathbf{B}^o_i(f \rightarrow g))$
6. $\models^{\mathrm{I}} (\neg f \wedge \langle \mathrm{do}_i(\mathtt{observe}\, f)\rangle \mathbf{B}^o_i g) \leftrightarrow (\neg f \wedge \mathbf{B}^o_i(\neg f \rightarrow g))$
7. $\models^{\mathrm{I}} \langle \mathrm{do}_i(\mathtt{observe}\, f)\rangle \varphi \leftrightarrow \langle \mathrm{do}_i(\mathtt{observe}\, \neg f)\rangle \varphi$
8. $\models^{\mathrm{I}} f \wedge \mathbf{Ignorant}^k_i f \rightarrow \langle \mathrm{do}_i(\mathtt{observe}\, f)\rangle \mathbf{Saw}_i f$
9. $\models^{\mathrm{I}} \neg f \wedge \mathbf{Ignorant}^k_i f \rightarrow \langle \mathrm{do}_i(\mathtt{observe}\, f)\rangle \mathbf{Saw}_i \neg f$
10. $\models^{\mathrm{I}} f \wedge (\mathbf{Heard}_i \neg f \vee \mathbf{Jumped}_i \neg f) \rightarrow \langle \mathrm{do}_i(\mathtt{observe}\, f)\rangle \mathbf{Saw}_i f$
11. $\models^{\mathrm{I}} f \wedge \mathbf{B}^c_i \neg f \rightarrow \langle \mathrm{do}_i(\mathtt{observe}\, f)\rangle ((\mathbf{B}^c_i \varphi \leftrightarrow \mathbf{B}^o_i \varphi) \wedge (\mathbf{B}^d_i \varphi \leftrightarrow \mathbf{B}^c_i \varphi))$

Proof: Let $\mathrm{M} \in \mathbf{M}^{\mathrm{I}}$ with state s, $f, g \in \mathrm{L}_0$, $i, j \in \mathrm{A}$ and $\varphi \in \mathrm{L}^{\mathrm{I}}$ be arbitrary.

1. Both o-informativeness and truthfulness are easily shown by using Proposition 5.24 and Proposition 5.26, respectively. We prove here that $\mathtt{observe}\, f$ is o-informative. The proof of this claim proceeds according to an established pattern, according to which also proofs that other informative actions satisfy other properties proceed.

   $\mathtt{observe}\, f$ is o-informative
   $\Leftrightarrow \mathbf{F}^{\mathrm{I}} \models^{\mathrm{I}} [\mathrm{do}_j(\mathtt{observe}\, f)]\mathbf{Bwhether}^o_j f$ as a schema in $j \in \mathrm{A}$
   $\Leftrightarrow \forall \mathrm{F} \in \mathbf{F}^{\mathrm{I}}(\mathrm{F} \models^{\mathrm{I}} [\mathrm{do}_j(\mathtt{observe}\, f)]\mathbf{Bwhether}^o_j f)$ as a schema in $j \in \mathrm{A}$
   $\Leftrightarrow \forall \mathrm{F} \in \mathbf{F}^{\mathrm{I}} \forall i \in \mathrm{A}(\mathrm{F} \models^{\mathrm{I}} [\mathrm{do}_i(\mathtt{observe}\, f)]\mathbf{Bwhether}^o_i f)$
   $\Leftrightarrow \forall \mathrm{F} \in \mathbf{F}^{\mathrm{I}} \forall i \in \mathrm{A} \forall \pi \in (\Pi \times \mathrm{S}) \rightarrow \mathsf{bool}((\mathrm{F}, \pi) \models^{\mathrm{I}} [\mathrm{do}_i(\mathtt{observe}\, f)]\mathbf{Bwhether}^o_i f)$
   $\Leftrightarrow \forall \mathrm{F} \in \mathbf{F}^{\mathrm{I}} \forall i \in \mathrm{A} \forall \pi \in (\Pi \times \mathrm{S}) \rightarrow \mathsf{bool} \forall s \in \mathrm{S}$
   $((\mathrm{F}, \pi), s \models^{\mathrm{I}} [\mathrm{do}_i(\mathtt{observe}\, f)]\mathbf{Bwhether}^o_i f)$

   Now let $\mathrm{F} \in \mathbf{F}^{\mathrm{I}}$, $i \in \mathrm{A}$, $\pi \in (\Pi \times \mathrm{S}) \rightarrow \mathsf{bool}$ and $s \in \mathrm{S}$ be arbitrary. Let furthermore $\mathrm{M}', s = \mathtt{r}^{\mathrm{I}}(i, \mathtt{observe}\, f)((\mathrm{F}, \pi), s)$; M' exists by Proposition 5.26. To show: $\mathrm{M}', s \models^{\mathrm{I}} \mathbf{Bwhether}^o_i f$. If $(\mathrm{F}, \pi), s \models^{\mathrm{I}} \mathbf{Bwhether}^o_i f$, then $\mathrm{M}' = (\mathrm{F}, \pi)$, and thus $\mathrm{M}', s \models^{\mathrm{I}} \mathbf{Bwhether}^o_i f$. If $(\mathrm{F}, \pi), s \models^{\mathrm{I}} f \wedge \mathbf{Ignorant}^o_i f$, then since $\mathrm{M}' = \mathtt{revise}^o(i, f)((\mathrm{F}, \pi), s)$, it follows by the first item of Proposition 5.24 that $\mathrm{M}', s \models^{\mathrm{I}}$

$\mathbf{B}_i^o f$, and thus $\mathrm{M}', s \models^{\mathrm{I}} \mathbf{Bwhether}_i^o f$. If $(\mathrm{F}, \pi), s \models^{\mathrm{I}} \neg f \wedge \mathbf{Ignorant}_i^o f$, then since $\mathrm{M}' = \mathtt{revise}^o(i, f)((\mathrm{F}, \pi), s)$, it follows by the second item of Proposition 5.24 that $\mathrm{M}', s \models^{\mathrm{I}} \mathbf{B}_i^o \neg f$, and thus $\mathrm{M}', s \models^{\mathrm{I}} \mathbf{Bwhether}_i^o f$. In all three cases it holds that $\mathrm{M}', s \models^{\mathrm{I}} \mathbf{Bwhether}_i^o f$, which suffices to conclude that $\mathtt{observe}\ f$ is o-informative.

2. Determinism and realisability follow directly from Proposition 5.26. Idempotence is easily shown by inspection of Definition 5.19.
3. Since observations leave the epistemic accessibility relation in a model intact, it follows that $\models^{\mathrm{I}} \mathbf{K}_j g \leftrightarrow [\mathrm{do}_i(\mathtt{observe}\ f)]\mathbf{K}_j g$. Due to the realisability and determinism of the observe action the box may be replaced by a diamond.
4. Let $\mathrm{M}', s = \mathtt{r}^{\mathrm{I}}(i, \mathtt{observe}\ f)(\mathrm{M}, s)$; M' exists since $\mathtt{observe}\ f$ is realisable. We distinguish two cases:

   $j \neq i$: In this case $\mathrm{B}^{o\prime}(j) = \mathrm{B}^o(j)$. Now $\mathrm{M}, s \models^{\mathrm{I}} \mathbf{B}_j^o g$ iff $\mathrm{M}, s' \models^{\mathrm{I}} g$ for all $(s, s') \in \mathrm{B}^o(j)$. Since $g \in \mathrm{L}_0$ we have that $\mathrm{M}, s' \models^{\mathrm{I}} g$ for all $(s, s') \in \mathrm{B}^o(j)$ iff $\mathrm{M}', s' \models^{\mathrm{I}} g$ for all $(s, s') \in \mathrm{B}^{o\prime}(j)$. Hence $\mathrm{M}, s \models^{\mathrm{I}} \mathbf{B}_j^o g \leftrightarrow \langle \mathrm{do}_i(\mathtt{observe}\ f)\rangle \mathbf{B}_j^o g$ in the case that $j \neq i$.

   $j = i$: From Proposition 5.24 it follows that $\mathrm{B}_o(i, \mathrm{M}, s) \subseteq \mathrm{B}_o(i, \mathrm{M}', s)$. Hence whenever $\mathrm{M}, s \models^{\mathrm{I}} \mathbf{B}_i^o g$ it follows that $\mathrm{M}', s \models^{\mathrm{I}} \mathbf{B}_i^o g$, which suffices to conclude that $\mathrm{M}, s \models^{\mathrm{I}} \mathbf{B}_i^o g \rightarrow \langle \mathrm{do}_i(\mathtt{observe}\ f)\rangle \mathbf{B}_i^o g$.

   Since in both cases $\mathrm{M}, s \models^{\mathrm{I}} \mathbf{B}_j^o g \rightarrow \langle \mathrm{do}_i(\mathtt{observe}\ f)\rangle \mathbf{B}_j^o g$ we conclude that item 4 holds.
5. Let $\mathrm{M}, s \models^{\mathrm{I}} f \wedge \langle \mathrm{do}_i(\mathtt{observe}\ f)\rangle \mathbf{B}_i^o g$ and let $\mathrm{M}' = \mathtt{r}^{\mathrm{I}}(i, \mathtt{observe}\ f)(\mathrm{M}, s)$. We distinguish two cases:
   - If $\mathrm{M}, s \models^{\mathrm{I}} \mathbf{Bwhether}_i^o f$, then $\mathrm{M}' = \mathrm{M}$ and $\mathrm{M}, s \models^{\mathrm{I}} f \wedge \langle \mathrm{do}_i(\mathtt{observe}\ f)\rangle \mathbf{B}_i^o g$ implies that $\mathrm{M}, s \models^{\mathrm{I}} \mathbf{B}_i^o g$, and thus $\mathrm{M}, s \models^{\mathrm{I}} \mathbf{B}_i^o(f \rightarrow g)$. Also, if $\mathrm{M}, s \models^{\mathrm{I}} f \wedge \mathbf{B}_i^o(f \rightarrow g)$ while $\mathrm{M}, s \models^{\mathrm{I}} \mathbf{Bwhether}_i^o f$, then $\mathrm{M}, s \models^{\mathrm{I}} \mathbf{B}_i^o f \wedge \mathbf{B}_i^o(f \rightarrow g)$, and thus $\mathrm{M}, s \models^{\mathrm{I}} \mathbf{B}_i^o g$. Hence $\mathrm{M}, s \models^{\mathrm{I}} \langle \mathrm{do}_i(\mathtt{observe}\ f)\rangle \mathbf{B}_i^o g$ and thus $\mathrm{M}, s \models^{\mathrm{I}} (f \wedge \langle \mathrm{do}_i(\mathtt{observe}\ f)\rangle \mathbf{B}_i^o g) \leftrightarrow (f \wedge \mathbf{B}_i^o(f \rightarrow g))$.
   - If $\mathrm{M}, s \models^{\mathrm{I}} \mathbf{Ignorant}_i^o f$, then $\mathrm{M}, s \models^{\mathrm{I}} f \wedge \langle \mathrm{do}_i(\mathtt{observe}\ f)\rangle \mathbf{B}_i^o g$ implies by similar arguments as given in the proof of Proposition 5.23 that $\mathrm{M}, s \models^{\mathrm{I}} \mathbf{B}_i^o(f \rightarrow g)$, and thus $\mathrm{M}, s \models^{\mathrm{I}} f \wedge \mathbf{B}_i^o(f \rightarrow g)$. Furthermore, if $\mathrm{M}, s \models^{\mathrm{I}} f \wedge \mathbf{B}_i^o(f \rightarrow g) \wedge \mathbf{Ignorant}_i^o f$, then $g \in \mathrm{B}_o^{\star f}(i, \mathrm{M}, s)$ — by Proposition 5.24(1) — hence $\mathrm{M}', s \models^{\mathrm{I}} \mathbf{B}_i^o g$. Thus $\mathrm{M}, s \models^{\mathrm{I}} f \wedge \langle \mathrm{do}_i(\mathtt{observe}\ f)\rangle \mathbf{B}_i^o g$, and also in this case $\mathrm{M}, s \models^{\mathrm{I}} (f \wedge \langle \mathrm{do}_i(\mathtt{observe}\ f)\rangle \mathbf{B}_i^o g) \leftrightarrow (f \wedge \mathbf{B}_i^o(f \rightarrow g))$.

   Since in both cases $\mathrm{M}, s \models^{\mathrm{I}} (f \wedge \langle \mathrm{do}_i(\mathtt{observe}\ f)\rangle \mathbf{B}_i^o g) \leftrightarrow (f \wedge \mathbf{B}_i^o(f \rightarrow g))$ it follows that item 5 indeed holds.
6. The proof of this item is completely analogous to the previous one.
7. Since $\models^{\mathrm{I}} \mathbf{Bwhether}_i^o f \leftrightarrow \mathbf{Bwhether}_i^o \neg f$ and furthermore $\mathtt{revise}^o(i, f)(\mathrm{M}, s) = \mathtt{revise}^o(i, \neg f)(\mathrm{M}, s)$ it follows that $\mathtt{r}^{\mathrm{I}}(i, \mathtt{observe}\ f)(\mathrm{M}, s)$ and $\mathtt{r}^{\mathrm{I}}(i, \mathtt{observe}\ \neg f)(\mathrm{M}, s)$ are identical.

8. Combining item 5 with the o-informativeness of $\texttt{observe}\, f$ with regard to f yields that if $\mathrm{M}, s \models^{\mathrm{I}} f \wedge \mathbf{Ignorant}_i^o f$, then $\mathrm{M}, s \models^{\mathrm{I}} \langle \mathrm{do}_i(\texttt{observe}\, f)\rangle \mathbf{B}_i^o f$. Since $\models^{\mathrm{I}} \mathbf{Ignorant}_i^o f \rightarrow \mathbf{Ignorant}_i^k f$, we have by item 3 that $\mathrm{M}, s \models^{\mathrm{I}} \mathbf{Ignorant}_i^k f \rightarrow \langle \mathrm{do}_i(\texttt{observe}\, f)\rangle \mathbf{Ignorant}_i^k f$, which suffices to conclude that this item indeed holds.
9. The proof of this item is completely analogous to the previous one.
10. If $\mathrm{M}, s \models^{\mathrm{I}} f \wedge (\mathbf{Heard}_i \neg f \vee \mathbf{Jumped}_i \neg f)$ then also $\mathrm{M}, s \models^{\mathrm{I}} f \wedge \mathbf{Ignorant}_i^k f$. From item 7 it then follows that $\mathrm{M}, s \models^{\mathrm{I}} \langle \mathrm{do}_i(\texttt{observe}\, f)\rangle \mathbf{Saw}_i f$.
11. Suppose $\mathrm{M}, s \models^{\mathrm{I}} f \wedge \mathbf{B}_i^c \neg f$. Then obviously $\mathrm{M}, s \models^{\mathrm{I}} f \wedge \mathbf{Ignorant}_i^o f$. Let $\mathrm{M}', s = \mathrm{r}^{\mathrm{I}}(i, \texttt{observe}\, f)(\mathrm{M}, s)$. By definition of $\texttt{revise}^o(i, f)$ it follows that
    - $[s]_{\mathrm{B}^{o\prime}(i)} = [s]_{\mathrm{B}^o(i)} \cap [\![f]\!]$
    - $\mathrm{B}^{c\prime}(i, s) = [s]_{\mathrm{B}^{o\prime}(i)}$
    - $\mathrm{B}^{d\prime}(i, s) = [s]_{\mathrm{B}^{o\prime}(i)}$

    Hence $\mathrm{M}', s \models^{\mathrm{I}} \mathbf{B}_i^o \varphi$ iff $\mathrm{M}', s \models^{\mathrm{I}} \mathbf{B}_i^c \varphi$ and $\mathrm{M}', s \models^{\mathrm{I}} \mathbf{B}_i^c \varphi$ iff $\mathrm{M}', s \models^{\mathrm{I}} \mathbf{B}_i^d \varphi$ which suffices to conclude item 10.

⊠

5.30. Proposition. *For all $i, i', j \in \mathrm{A}$, $f, g \in \mathrm{L}_0$ and $\varphi \in \mathrm{L}^{\mathrm{I}}$, we have:*

1. $\texttt{inform}\,(f, i)$ *is deterministic and idempotent*
2. $\models^{\mathrm{I}} \mathbf{B}_{i'}^x g \rightarrow [\mathrm{do}_j(\texttt{inform}\,(f, i))] \mathbf{B}_{i'}^x g$ *for* $x \in \{k, o\}$
3. $\models^{\mathrm{I}} \mathbf{B}_j^d f \leftrightarrow \langle \mathrm{do}_j(\texttt{inform}\,(f, i))\rangle \top$
4. $\models^{\mathrm{I}} \mathbf{B}_j^d f \wedge \neg \mathbf{D}_{i,j} f \rightarrow (\langle \mathrm{do}_j(\texttt{inform}\,(f, i))\rangle \varphi \leftrightarrow \varphi)$
5. $\models^{\mathrm{I}} \mathbf{D}_{i,j} f \wedge \mathbf{B}_j^c f \rightarrow \langle \mathrm{do}_j(\texttt{inform}\,(f, i))\rangle \mathbf{Bwhether}_i^c f$
6. $\models^{\mathrm{I}} \mathbf{D}_{i,j} f \wedge \mathbf{B}_j^c f \wedge \mathbf{Ignorant}_i^c f \rightarrow \langle \mathrm{do}_j(\texttt{inform}\,(f, i))\rangle \mathbf{Heard}_i f$
7. $\models^{\mathrm{I}} \mathbf{D}_{i,j} f \wedge \mathbf{Heard}_j f \wedge \mathbf{Bwhether}_i^c f \rightarrow (\langle \mathrm{do}_j(\texttt{inform}\,(f, i))\rangle \varphi \leftrightarrow \varphi)$
8. $\models^{\mathrm{I}} \mathbf{D}_{i,j} f \wedge \mathbf{B}_j^c f \wedge \mathbf{Ignorant}_i^c f \rightarrow (\langle \mathrm{do}_j(\texttt{inform}\,(f, i))\rangle \mathbf{B}_i^c g \leftrightarrow \mathbf{B}_i^c (f \rightarrow g))$
9. $\models^{\mathrm{I}} \mathbf{D}_{i,j} f \wedge \mathbf{B}_j^o f \wedge \mathbf{Ignorant}_i^o f \wedge \mathbf{B}_i^c \neg f \rightarrow (\langle \mathrm{do}_j(\texttt{inform}\,(f, i))\rangle \mathbf{B}_i^c g \leftrightarrow \mathbf{B}_i^o (f \rightarrow g))$
10. $\models^{\mathrm{I}} \mathbf{D}_{i,j} f \wedge \mathbf{Jumped}_j f \rightarrow (\langle \mathrm{do}_j(\texttt{inform}\,(f, i))\rangle \varphi \leftrightarrow \varphi)$
11. $\models^{\mathrm{I}} \mathbf{B}_i^x f \rightarrow (\langle \mathrm{do}_i(\texttt{inform}\, f, i))\rangle \varphi \leftrightarrow \varphi)$ *for* $x \in \{k, o, c, d\}$

Proof: Let $\mathrm{M} \in \mathbf{M}^{\mathrm{I}}$ with state s, $i, j \in \mathrm{A}$, $f, g \in \mathrm{L}_0$ and $\varphi \in \mathrm{L}^{\mathrm{I}}$ be arbitrary.

1. Determinism and idempotence of $\texttt{inform}\,(f, i)$ are easily shown by inspection of Definitions 5.19 and 5.28.
2. This item follows since communication affects neither the epistemic accessibility relation nor the observational belief clusters of any agent.
3. This item is straightforward from Definition 5.28 and Proposition 5.20(2).
4. If $\mathrm{M}, s \models^{\mathrm{I}} \mathbf{B}_j^d f \wedge \neg \mathbf{D}_{i,j} f$ then, by Definition 5.28, $\mathrm{r}^{\mathrm{I}}(j, \texttt{inform}\,(f, i))(\mathrm{M}, s) = \mathrm{M}, s$. Hence $\mathrm{M}, s \models^{\mathrm{I}} \langle \mathrm{do}_j(\texttt{inform}\,(f, i))\rangle \varphi$ iff $\mathrm{M}, s \models^{\mathrm{I}} \varphi$.
5. Assume that $\mathrm{M}, s \models^{\mathrm{I}} \mathbf{D}_{i,j} f \wedge \mathbf{B}_j^c f$. By item 3 we know that M', s exists such that $\mathrm{M}', s = \mathrm{r}^{\mathrm{I}}(j, \texttt{inform}\,(f, i))(\mathrm{M}, s)$. We distinguish two cases:

- $\mathrm{M}, s \models^{\mathrm{I}} (\mathbf{B}^o_j f \wedge \mathbf{Ignorant}^o_i f) \vee \mathbf{Ignorant}^c_i f$. In this case $\mathrm{M}' = \mathtt{revise}^c(i, f)(\mathrm{M}, s)$. By Proposition 5.23(1) it then follows that $\mathrm{M}', s \models^{\mathrm{I}} \mathbf{B}^c_i f$, and hence $\mathrm{M}', s \models^{\mathrm{I}} \mathbf{Bwhether}^c_i f$, which was to be shown.
- $\mathrm{M}, s \not\models^{\mathrm{I}} (\mathbf{B}^o_j f \wedge \mathbf{Ignorant}^o_i f) \vee \mathbf{Ignorant}^c_i f$. In this case $\mathrm{M}' = \mathrm{M}$, so we have to show that $\mathrm{M}, s \models^{\mathrm{I}} \mathbf{Bwhether}^c_i f$. We have that $\mathrm{M}, s \models^{\mathrm{I}} (\neg \mathbf{B}^o_i f \vee \neg \mathbf{Ignorant}^o_i f) \wedge \neg \mathbf{Ignorant}^c_i f$, which implies $\mathrm{M}, s \models^{\mathrm{I}} \neg \mathbf{Ignorant}^c_i f$, and $\mathrm{M}, s \models^{\mathrm{I}} \mathbf{Bwhether}^c_i f$.

In both cases $\mathrm{M}', s \models^{\mathrm{I}} \mathbf{Bwhether}^c_i f$. Thus $\mathrm{M}, s \models^{\mathrm{I}} \langle \mathrm{do}_j(\mathtt{inform}\,(f, i))\rangle \mathbf{Bwhether}^c_i f$, which suffices to conclude item 5.

6. Assume that $\mathrm{M}, s \models^{\mathrm{I}} \mathbf{D}_{i,j} f \wedge \mathbf{B}^c_j f \wedge \mathbf{Ignorant}^c_i f$. Let M' be the model such that $\mathrm{M}', s = \mathtt{r}^{\mathrm{I}}(j, \mathtt{inform}\,(f, i))(\mathrm{M}, s)$. By clause 2 of Definition 5.28 it follows that $\mathrm{M}' = \mathtt{revise}^c(i, f)(\mathrm{M}, s)$. By Proposition 5.23(1) we have that $\mathrm{M}', s \models^{\mathrm{I}} \mathbf{B}^c_i f$. Since $\mathrm{M}, s \models^{\mathrm{I}} \mathbf{Ignorant}^c_i f$ implies $\mathrm{M}, s \models^{\mathrm{I}} \mathbf{Ignorant}^o_i f$, it follows that $\mathrm{M}, s \models^{\mathrm{I}} \mathbf{Ignorant}^o_i f$, and since communication does not affect observational beliefs on propositional formulae, we have that $\mathrm{M}', s \models^{\mathrm{I}} \mathbf{Ignorant}^o_i f$. Hence $\mathrm{M}', s \models^{\mathrm{I}} \mathbf{Heard}_i f$ and $\mathrm{M}, s \models^{\mathrm{I}} \langle \mathrm{do}_j(\mathtt{inform}\,(f, i))\rangle \mathbf{Heard}_i f$.
7. If $\mathrm{M}, s \models^{\mathrm{I}} \mathbf{D}_{i,j} f \wedge \mathbf{Heard}_j f \wedge \mathbf{Bwhether}^c_i f$, then $\mathrm{M}, s \not\models^{\mathrm{I}} \neg \mathbf{B}^d_j f$, $\mathrm{M}, s \not\models^{\mathrm{I}} (\mathbf{B}^o_j f \wedge \mathbf{Ignorant}^o_i f)$ and $\mathrm{M}, s \not\models^{\mathrm{I}} (\mathbf{Heard}_j f \wedge \mathbf{Ignorant}^c_i f)$, which implies that $\mathtt{r}^{\mathrm{I}}(j, \mathtt{inform}\,(f, i))(\mathrm{M}, s) = \mathrm{M}, s$. Thus $\mathrm{M}, s \models^{\mathrm{I}} \varphi$ iff $\mathrm{M}, s \models^{\mathrm{I}} \langle \mathrm{do}_j(\mathtt{inform}\,(f, i))\rangle \varphi$, which suffices to conclude that item 7 holds.
8. The proof of this item proceeds along the lines of the ones given for items 5 and 6 of Proposition 5.27.
9. Let $\mathrm{M}, s \models^{\mathrm{I}} \mathbf{D}_{i,j} f \wedge \mathbf{B}^o_j f \wedge \mathbf{Ignorant}^o_i f \wedge \mathbf{B}^c_i \neg f$. Let $\mathrm{M}', s = \mathtt{r}^{\mathrm{I}}(j, \mathtt{inform}\,(f, i))(\mathrm{M}, s)$; M' exists by item 3. Inspection of Definition 5.28 shows that $\mathrm{M}' = \mathtt{revise}^c(i, f)(\mathrm{M}, s)$. Inspection of Definition 5.19 shows on its turn that $\mathrm{B}^{c\prime}(i, s) = [s]_{\mathrm{B}^o(i)} \cap [\![f]\!]$. By similar arguments as given above one concludes that $\mathrm{M}', s \models^{\mathrm{I}} \mathbf{B}^c_i g$ iff $\mathrm{M}, s \models^{\mathrm{I}} \mathbf{B}^o_i (f \rightarrow g)$, which suffices to conclude that item 9 holds.
10. This item is straightforward from Definition 5.28: if $\mathrm{M}, s \models^{\mathrm{I}} \mathbf{D}_{i,j} f \wedge \mathbf{Jumped}_j f$ it follows that $\mathtt{r}^{\mathrm{I}}(j, \mathtt{inform}\,(f, i))(\mathrm{M}, s) = \mathrm{M}, s$ and hence $\mathrm{M}, s \models^{\mathrm{I}} \langle \mathrm{do}_j(\mathtt{inform}\,(f, i))\rangle \varphi$ iff $\mathrm{M}, s \models^{\mathrm{I}} \varphi$.
11. Let $x \in \{d, c, o, k\}$ be arbitrary, and assume that $\mathrm{M}, s \models^{\mathrm{I}} \mathbf{B}^x_i f$. By item 3 it follows that $\mathrm{M}', s = \mathtt{r}^{\mathrm{I}}(i, \mathtt{inform}\,(f, i))(\mathrm{M}, s)$ exists. Since both $\mathbf{B}^o_i f \wedge \mathbf{Ignorant}^o_i f$ and $\mathbf{Heard}_i f \wedge \mathbf{Ignorant}^c_i f$ are not satisfiable, and therefore not true in s, the second clause of Definition 5.28 is not applicable, and hence $\mathrm{M}' = \mathrm{M}$. Hence $\mathrm{M}, s \models^{\mathrm{I}} \varphi$ iff $\mathrm{M}', s \models^{\mathrm{I}} \varphi$, which concludes the proof of item 11.

⊠

5.36. Proposition. *For all $i, j \in \mathrm{A}$, $f, g \in \mathrm{L}_0$ and $\varphi \in \mathrm{L}^{\mathrm{I}}$, we have:*

1. $\mathtt{try\_jump}\, f$ *is deterministic, idempotent and d-informative with regard to f*
2. $\models^{\mathrm{I}} \mathbf{B}^x_j g \rightarrow [\mathrm{do}_i(\mathtt{try\_jump}\, f)]\mathbf{B}^x_j g$ *for $x \in \{d, c, o, k\}$*

3. $\models^{\mathrm{I}} \mathbf{N}f \leftrightarrow \langle \mathrm{do}_i(\mathtt{try\_jump}\, f)\rangle \top$
4. $\models^{\mathrm{I}} \langle \mathrm{do}_i(\mathtt{try\_jump}\, f)\rangle \top \leftrightarrow \langle \mathrm{do}_i(\mathtt{try\_jump}\, f)\rangle \mathbf{Bwhether}_i^d f$
5. $\models^{\mathrm{I}} \mathbf{N}f \wedge \mathbf{Ignorant}_i^d f \rightarrow \langle \mathrm{do}_i(\mathtt{try\_jump}\, f)\rangle \mathbf{Jumped}_i f$
6. $\models^{\mathrm{I}} \mathbf{N}f \wedge \mathbf{Ignorant}_i^d f \rightarrow (\langle \mathrm{do}_i(\mathtt{try\_jump}\, f)\rangle \mathbf{B}_i^d g \leftrightarrow \mathbf{B}_i^d(f \rightarrow g))$
7. $\models^{\mathrm{I}} \mathbf{N}f \wedge \mathbf{Bwhether}_i^d f \rightarrow (\langle \mathrm{do}_i(\mathtt{try\_jump}\, f)\rangle \varphi \leftrightarrow \varphi)$

Proof: Let $\mathrm{M} \in \mathbf{M}^{\mathrm{I}}$ with state s, $i, j \in \mathrm{A}$, $f, g \in \mathrm{L}_0$ and $\varphi \in \mathrm{L}^{\mathrm{I}}$ be arbitrary.

1. Determinism and idempotence follow by inspection of Definitions 5.19 and 5.34. Informativeness is shown as it was for the $\mathtt{observe}\, f$ actions in item 1 of Proposition 5.27.
2. In the case that $\mathrm{r}^{\mathrm{I}}(i, \mathtt{try\_jump}\, f)(\mathrm{M}, s) = \emptyset$, the proposition trivially holds. Hence assume that $\mathrm{r}^{\mathrm{I}}(i, \mathtt{try\_jump}\, f)(\mathrm{M}, s) = \mathrm{M}'$. In this case one concludes by a similar argument as given in the proof of item 4 of Proposition 5.27 that $\mathrm{M}, s \models^{\mathrm{I}} \mathbf{B}_j^x g \leftrightarrow [\mathrm{do}_i(\mathtt{try\_jump}\, f)]\mathbf{B}_j^x g$ in the case that $j \neq i$ or $x \in \{c, o, k\}$ and $\mathrm{M}, s \models^{\mathrm{I}} \mathbf{B}_i^d g \rightarrow [\mathrm{do}_i(\mathtt{try\_jump}\, f)]\mathbf{B}_i^d g$.
3. If $\mathrm{M}, s \models^{\mathrm{I}} \mathbf{N}f$, M' exists such that $\mathrm{M}', s = \mathrm{r}^{\mathrm{I}}(i, \mathtt{try\_jump}\, f)(\mathrm{M}, s)$: if $\mathrm{M}, s \not\models^{\mathrm{I}} \mathbf{Ignorant}_i^d f$, $\mathrm{M}' = \mathrm{M}$ and else M' exists by definition of the $\mathtt{revise}^d$ function and Proposition 5.20(2). The reverse implication is even so obvious.
4. Suppose $\mathrm{M}, s \models^{\mathrm{I}} \langle \mathrm{do}_i(\mathtt{try\_jump}\, f)\rangle \top$, and let $\mathrm{M}', s = \mathrm{r}^{\mathrm{I}}(i, \mathtt{try\_jump}\, f)(\mathrm{M}, s)$. We distinguish two cases:
   - If $\mathrm{M}, s \not\models^{\mathrm{I}} \mathbf{Ignorant}_i^d f$ then $\mathrm{M}' = \mathrm{M}$. Hence $\mathrm{M}', s \models^{\mathrm{I}} \neg\mathbf{Ignorant}_i^d f$ and thus $\mathrm{M}, s \models^{\mathrm{I}} \langle \mathrm{do}_i(\mathtt{try\_jump}\, \varphi)\rangle \mathbf{Bwhether}_i^d \varphi$.
   - If $\mathrm{M}, s \models^{\mathrm{I}} \mathbf{Ignorant}_i^d \varphi$ then $\mathrm{M} = \mathtt{revise}^d(i, f)(\mathrm{M}, s)$. By Proposition 5.23(1) it follows that $\mathrm{M}', s \models^{\mathrm{I}} \mathbf{B}_i^d f$ and hence $\mathrm{M}, s \models^{\mathrm{I}} \langle \mathrm{do}_i(\mathtt{try\_jump}\, f)\rangle \mathbf{Bwhether}_i^d f$.

   Since in both cases $\mathrm{M}, s \models^{\mathrm{I}} \langle \mathrm{do}_i(\mathtt{try\_jump}\, f)\rangle \mathbf{Bwhether}_i^d f$ we conclude that the left-to-right implication of this item holds. The right-to-left implication is trivial.
5. Suppose $\mathrm{M}, s \models^{\mathrm{I}} \mathbf{N}f \wedge \mathbf{Ignorant}_i^d f$. Let $\mathrm{M}', s = \mathrm{r}^{\mathrm{I}}(i, \mathtt{try\_jump}\, f)(\mathrm{M}, s)$. From Definition 5.34 it follows that $\mathrm{M}' = \mathtt{revise}^d(i, f)(\mathrm{M}, s)$. By Proposition 5.23(1) it follows that $\mathrm{M}', s \models^{\mathrm{I}} \mathbf{B}_i^d f$. From $\mathrm{M}, s \models^{\mathrm{I}} \mathbf{Ignorant}_i^d f$ it follows that $\mathrm{M}, s \models^{\mathrm{I}} \mathbf{Ignorant}_i^c f$. Since attempted jumps to conclusions do not affect the communicational beliefs of agents on propositional formulae, it follows that $\mathrm{M}', s \models^{\mathrm{I}} \mathbf{Ignorant}_i^c f$ and thus $\mathrm{M}', s \models^{\mathrm{I}} \mathbf{Jumped}_i f$. Then also $\mathrm{M}, s \models^{\mathrm{I}} \langle \mathrm{do}_i(\mathtt{try\_jump}\, f)\rangle \mathbf{Jumped}_i f$.
6. This item is shown in similar ways as item 8 of Proposition 5.30 and items 5 and 6 of Proposition 5.27.
7. Suppose $\mathrm{M}, s \models^{\mathrm{I}} \mathbf{N}f \wedge \mathbf{Bwhether}_i^d f$. Then it follows from Definition 5.34 that $\mathrm{r}^{\mathrm{I}}(i, \mathtt{try\_jump}\, f)(\mathrm{M}, s) = \mathrm{M}, s$. But then $\mathrm{M}, s \models^{\mathrm{I}} \langle \mathrm{do}_i(\mathtt{try\_jump}\, f)\rangle \varphi$ iff $\mathrm{M}, s \models^{\mathrm{I}} \varphi$.

⊠

Chapter 6

How to motivate your agents

> *Wir stellen uns die Frage, ob an der Arbeit unseres seelischen Apparates eine Hauptabsicht zu erkennen sei, und beantworten sie in erster Annäherung, daß diese Absicht auf Lustgewinnung gerichtet ist. Es scheint, daß unsere gesamte Seelentätigkeit darauf gerichtet ist, Lust zu erwerben und Unlust zu vermeiden, daß sie automatisch durch das Lust-prinzip reguliert wird. Nun wüßten wir um alles in der Welt gerne, welches die Bedingungen der Enstehung von Lust und Unlust sein, aber daran fehlt es uns eben.*
>
> Sigmund Freud, *'Vorlesungen zur Einführung in die Psychoanalyse'*.

In this chapter we present a formalisation of motivational attitudes, the attitudes that explain why agents act. We consider the statics of these attitudes both at the assertion level, i.e. ranging over propositions, and at the practition[1] level, i.e. ranging over actions, as well as the dynamics of these attitudes. Starting from an agent's wishes, which form the primitive, most fundamental motivational attitude, we define its goals as induced by those wishes that do not yet hold, i.e. are unfulfilled, but are within the agent's practical possibility to bring about, i.e. are implementable for the agent. Among these unfulfilled, implementable wishes the agent selects those that qualify as its goals. Based on its knowledge on its goals and practical possibilities, an agent may make certain commitments. In particular, an agent may commit itself to actions that it knows to be correct and feasible to bring about some of its known goals. As soon as it no longer knows its commitments to be useful, i.e. leading to fulfilment of some goal, and practically possible, an agent has the practical possibility to undo these commitments. Both the act of committing as well as that of undoing commitments is modelled as a special model-transforming action in our framework, according to the generalised paradigm for

[1] The term 'practition' is due to Castañeda [14].

Propositional Dynamic Logic as introduced in the previous chapter. In between making and undoing commitments, an agent is committed to all the actions that are for all practical purposes identical to the ones in its agenda. By recording finite computation runs of actions rather than the actions themselves in the agent's agenda, it is ensured that commitments display an acceptable behaviour with regard to composite actions. As usual, we conclude this chapter with a brief summary, some guidelines for future research, an overview of the relevant literature, and proofs of selected propositions.

6.1 Motivational attitudes: wishes, goals and commitments

Motivational attitudes constitute what probably is the most fundamental, primitive and essential characteristic of agency[2]. These attitudes provide the motive for any act on behalf of the agents, i.e. the acting of agents is driven by their motivational attitudes. Typical examples of motivational attitudes are amongst others wishes, desires, preferences, concerns, ambitions, goals, intentions and commitments. The meaning of most of these terms is intuitively much less clear than that of the informational attitudes of knowledge and belief, or of the aspects of action (result, opportunity, ability) that we considered. It is therefore also not clear which of the aforementioned motivational attitudes are relevant, and worth formalising, when modelling rational agents. In their BDI-architecture, Rao & Georgeff [109] consider desires and intentions to be primitive, and define a notion of commitment in terms of these, Cohen & Levesque [20] consider goals to be primitive and define intentions using goals, and Shoham [121] restricts himself to formalising commitments. In our opinion each of these formalisations lacks some of the aspects that are vital to modelling motivational attitudes. Firstly, psychological evidence seems to suggest that notions like goals, intentions and commitments are not primitive, but rather induced by some more fundamental notion. There is an ongoing debate in the psychological literature on the actual nature of this notion. Aristotle, and in his footsteps the adherents of cognitive motivation theories, proposed that the pursuit of knowledge is men's most primitive motivational attitude. Freud distinguished in his psychoanalytical theory two fundamental human motivations, viz. libido, the complex of desires associated with sexuality and the erotic, and aggression, which comprises besides a proclivity towards destruction also the propensity to self-preservation. Although the theories of Aristotle and Freud are readily applicable to human agents, the question arises whether this is also the case for non-human, artificial agents. We argue that this is indeed so. For highly advanced artificial agents, like for instance the robots appearing

[2]There are two main reasons for not dealing with these attitudes until this chapter, in spite of the fact that they are fundamental to agency. The first of these is the inherent complexity of formalisations of motivational attitudes. Related to this, an adequate modelling of motivational attitudes will in general necessitate a formalisation of most of the concepts considered in the previous chapters.

in the famous robot adventures by Asimov [4, 5, 6], are so much like humans that they can safely be ascribed human motives. Software agents that assist some user with some specific task, are in general not to be ascribed human motives. The primitive motivational attitudes of these agents could however quite well be identified with those of the user they are assisting. Lastly, the intelligent information agents that we considered in the previous chapter could be said to be driven by a quest for information, which would make them obey Aristotle's theory. In this chapter we will simply assume that all agents, human and artificial alike, posses some fundamental motivational attitude — whatever it may actually be.

Secondly, in addition to being faithful to insights gained in psychology, we feel that it is also important to pay attention to those gained in analytical philosophy. More specifically, we are of the opinion that the modelling of *practical reasoning* should be part of any formalisation of motivational attitudes that pretends to be an adequate one. The term 'practical reasoning' dates back to Aristotle, and refers to the process through which (human) agents conclude that they should perform certain actions in order to bring about some of the things that they like to be the case. It seems very likely that for autonomous agents in AI applications, which have to act (autonomously) to achieve some of their goals, practical reasoning accounts for the most essential and most frequently used kind of information processing. Hence an adequate formalisation of motivational attitudes should pay at least some attention to this kind of reasoning.

The third essential facet of any formalisation of motivational attitudes consists of the modelling of the act of *selecting*: agents have to make choices among the things they like to be the case, thereby deciding which of these they will try to achieve next. For it might be impossible to satisfy all of an agent's wishes simultaneously, since these wishes are either analytically inconsistent or incompatible given the agent's resources.

In our opinion all of the aspects mentioned above should be present in an adequate formalisation of motivational attitudes, and they are indeed so in the one that is presented in this chapter. The notions that are essential in this formalisation are *wishes*, *goals* and *commitments*. Of these, wishes constitute the primitive motivational attitude that models the things that an agent likes to be the case. As such, wishes naturally range over propositions, corresponding to the idea that agents wish for certain aspects of the world. We formalise wishes through a normal modal operator, i.e. an operator validating just the K-axiom and the N-rule. Agents set their goals by selecting among their wishes. However, agents are not allowed to select arbitrary wishes as their goals, but instead may only select wishes that are unfulfilled yet implementable. Whenever an agent knows that it has some goal, it may commit itself to any action that it knows to be correct and feasible with respect to the goal. This act of committing to an action is itself formalised as a special kind of action. Commitments to actions are in general to persist until all of the goals for which the commitment was made are fulfilled. Having said so, agents

should not be forced to remain committed to actions that have either become useless in that they do not lead to fulfilment of any goal, or impossible in that the agent no longer knows that it has the opportunity and ability to perform the action. Phrased differently, an agent should be allowed to uncommit itself whenever an action is no longer known to be correct and feasible with respect to one of the agent's goals.

To formalise wishes, goals and commitments and their associated concepts, we introduce a modal operator modelling wishes, operators modelling implementability, (made) choices and (made) commitments, and action constructors modelling the acts of selecting, committing and uncommitting.

6.1. Definition. To define the language L^{C}, the alphabet is extended with the wish operator $\mathbf{W}\_$, the implementability operator $\Diamond\_$, the selected operator $\mathbf{C}\_$, the commitment operator $\mathbf{Committed}\_$ and the action constructors `select` $\_$, `commit_to` $\_$ and `uncommit` $\_$.

The acts of committing and uncommitting are of an essentially different nature than the regular actions, execution of which changes the state of the world. Through the former actions agents (un)commit themselves to actions of the latter kind. Intuitively it does not make much sense to allow agents to commit themselves to making commitments: it is not at all clear how a statement like 'i is committed to commit itself to do α' is to be interpreted. Also statements like 'it is implementable for agent i to become committed' seem to be of a rather questionable nature. To avoid these kinds of counterintuitive situations, we define the language L^{C} on top of the language L as defined in Chapter 3. That is, the operators modelling wishes, implementability and selections are defined in such a way that they range over formulae from L rather than those from L^{C}. The operator modelling the commitments that an agent has made is defined to range over the actions from Ac, the class of actions associated with L, and not over Ac^{C}. Analogously, the special actions in Ac^{C}, as there are the action modelling the act of selecting and those modelling the making and undoing of commitments, range over elements from L and Ac rather than L^{C} and Ac^{C}.

6.2. Definition. The language L and the class Ac of actions are as in Definition 3.1, i.e. L is the smallest superset of Π closed under the core clauses and Ac is the smallest superset of At satisfying the core clauses.

The language L^{C} is the smallest superset of Π such that the core clauses are validated and furthermore

- if $\varphi \in \mathrm{L}$ and $i \in \mathrm{A}$ then $\mathbf{W}_i\varphi \in \mathrm{L}^{\mathrm{C}}$
- if $\varphi \in \mathrm{L}$ and $i \in \mathrm{A}$ then $\Diamond_i\varphi \in \mathrm{L}^{\mathrm{C}}$
- if $\varphi \in \mathrm{L}$ and $i \in \mathrm{A}$ then $\mathbf{C}_i\varphi \in \mathrm{L}^{\mathrm{C}}$
- if $\alpha \in \mathrm{Ac}$ and $i \in \mathrm{A}$ then $\mathbf{Committed}_i\alpha \in \mathrm{L}^{\mathrm{C}}$

The class Ac^{C} is the smallest superset of At closed under the core clauses and such that

- if $\varphi \in \mathrm{L}$ then $\mathtt{select}\,\varphi \in \mathrm{Ac}^{\mathrm{C}}$
- if $\alpha \in \mathrm{Ac}$ then $\mathtt{commit\_to}\,\alpha \in \mathrm{Ac}^{\mathrm{C}}$
- if $\alpha \in \mathrm{Ac}$ then $\mathtt{uncommit}\,\alpha \in \mathrm{Ac}^{\mathrm{C}}$

6.3. Definition. For $i \in \mathrm{A}$, $\alpha \in \mathrm{Ac}^{\mathrm{C}}$ and $\varphi \in \mathrm{L}^{\mathrm{C}}$, the abbreviations $\mathbf{Correct}_i(\alpha, \varphi)$, $\mathbf{Feasible}_i\alpha$, $\mathbf{PracPoss}_i(\alpha, \varphi)$ and $\mathbf{Can}_i(\alpha, \varphi)$ are defined as in the language L.

The models for the language L^{C} are equipped with elements used to interpret the agents' wishes, selections and commitments. Wishes are interpreted through an accessibility relation on worlds that denotes worlds that are more desirable from the agent's point of view. Selections are straightforwardly interpreted through a set of formulae that denotes the choices that an agent has made. From a formal point of view, this set acts as a kind of *awareness* on the wishes of an agent, thereby ensuring an intuitively acceptable behaviour of goals. Originally, Fagin & Halpern [31] introduced the idea of awareness sets as a means to solve the so-called problems of logical omniscience. As we will see in Section 6.3, the effect of the selection sets on the behaviour of goals is similar to that of the awareness sets on the properties of knowledge. The agents' commitments are interpreted by means of the agenda function, which yields for each agent in every state the commitments that it has made and is up to. Detailed accounts of the respective interpretations are given in the following sections.

6.4. Definition. A model M for the language L^{C} is a tuple containing the core elements, the functions $\mathrm{W} : \mathrm{A} \to \wp(\mathrm{S} \times \mathrm{S})$, which determines the desirability relation of an agent in a state, and $\mathrm{C} : \mathrm{A} \times \mathrm{S} \to \wp(\mathrm{L})$ denoting the choices made by an agent in a state, and a function $\mathrm{Agenda} : \mathrm{A} \times \mathrm{S} \to \wp(\mathrm{Ac_b})$, which records the commitments of agents.

As we did for the informative actions of Chapter 5, we interpret the acts of selecting, committing and uncommitting as model-transformations. Whereas the informative actions transformed models by modifying belief functions, the act of selecting does so by affecting the set of choices, and the act of (un)committing transforms the agent's agenda. To account for these modifications, we introduce the set of possible result models of a given model for L^{C} analogously to the set $\mathrm{M}^{\mathrm{I}}_{\sim}$ defined in the previous chapter.

6.5. Definition. Let $\mathrm{M} \in \mathbf{M}^{\mathrm{C}}$ be some model for L^{C}. The class $\mathrm{M}^{\mathrm{C}}_{\sim} \subseteq \mathbf{M}^{\mathrm{C}}$ contains all models that (possibly) differ from M only in the C or the Agenda functions.

The dynamic and ability formulae from L^{C} are interpreted as those from L^{I}, replacing $\models^{\mathrm{I}}$ by $\models^{\mathrm{C}}$ and $\mathrm{M}^{\mathrm{I}}_{\sim}$ by $\mathrm{M}^{\mathrm{C}}_{\sim}$. We will not repeat these definitions here (the reader is kindly referred to Chapter 5), but instead focus on the novel elements of the system of this chapter, the first of which is our formalisation of wishes.

6.2 Formalising wishes

Wishes are the most primitive, fundamental motivational attitudes, i.e. *in ultimo* agents are motivated to fulfil their wishes. As mentioned in Section 6.1, we formalise wishes through a plain normal modal operator, i.e. wishes are straightforwardly interpreted as a necessity operator over the accessibility relation W.

6.6. Definition. The binary relation $\models^{\mathrm{C}}$ between a formula in L^{C} and a pair M, s consisting of a model M for L^{C} and a state s in M is for wishes defined as follows:

$\mathrm{M}, s \models^{\mathrm{C}} \mathbf{W}_i\varphi \Leftrightarrow \forall s' \in \mathrm{S}((s, s') \in \mathrm{W}(i) \Rightarrow \mathrm{M}, s' \models^{\mathrm{C}} \varphi)$

It is well-known that normal modal operators have certain properties that are occasionally considered undesirable for the commonsense notions that they are intended to formalise. For example, although the formal notions of knowledge and belief are closed under logical consequence, this property will in general not hold for human knowledge and belief (although it will for instance hold for the information that is recorded in a database, or for the knowledge and belief of an artificial agent). When formalising motivational attitudes the undesired properties induced by closure under logical consequence become even more pregnant. For agents do in general not desire all the logical consequences of their wishes, nor do they consider the logically inevitable to be among their goals. For example, an agent that wants its teeth to be restored will in general not want or wish for the pain that inevitably accompanies such a restoration. And although the sun rises in the east there will hardly be an agent that desires this to be the case. The problem embodied by the former example is known as the *side-effect* problem; the problem that all logical tautologies are wishes (goals) of an agent is known as the *transference* problem. Both in syntactical shape as in meaning, these problems are closely related to the problems of logical omniscience that have plagued formalisations of informational attitudes for many years. In terms of our framework, seven of the most (in)famous problems of logical omniscience can be formulated as follows.

6.7. Definition. Let $\varphi, \psi \in \mathrm{L}^{\mathrm{X}}$ be formulae, and let $\mathbf{X}$ be some operator.

- $\models^{\mathrm{X}} \mathbf{X}\varphi \wedge \mathbf{X}(\varphi \rightarrow \psi) \rightarrow \mathbf{X}\psi$ LO1
- $\models^{\mathrm{X}} \varphi \Rightarrow \models^{\mathrm{X}} \mathbf{X}\varphi$ LO2
- $\models^{\mathrm{X}} \varphi \rightarrow \psi \Rightarrow \models^{\mathrm{X}} \mathbf{X}\varphi \rightarrow \mathbf{X}\psi$ LO3
- $\models^{\mathrm{X}} \varphi \leftrightarrow \psi \Rightarrow \models^{\mathrm{X}} \mathbf{X}\varphi \leftrightarrow \mathbf{X}\psi$ LO4
- $\models^{\mathrm{X}} (\mathbf{X}\varphi \wedge \mathbf{X}\psi) \rightarrow \mathbf{X}(\varphi \wedge \psi)$ LO5
- $\models^{\mathrm{X}} \mathbf{X}\varphi \rightarrow \mathbf{X}(\varphi \vee \psi)$ LO6
- $\models^{\mathrm{X}} \neg(\mathbf{X}\varphi \wedge \mathbf{X}\neg\varphi)$ LO7

Properties LO1 and LO3 as given in Definition 6.7 capture the side-effect problem, and property LO2 captures the transference problem. Of the other properties given not all are equally harmful when formalising wishes. In our opinion, property LO4 is not that harmful, and could even be considered desirable, dependent on the demands for rationality that one is willing to make. Property LO5, which we like to think of as representing 'the problem of *unrestricted combining*', is in general undesirable when formalising motivational attitudes. This is for instance shown by the example of an agent that likes watching TV and likes to read a book, while not wanting to watch TV and read a book at the same time. Property LO6, for which we coin the term 'the problem of *unrestricted weakening*', is a special instantiation of the side-effect problem. That this property is undesirable is shown by the example of an agent desiring itself to be painted green, without desiring being green or being crushed under a steam roller[3]. Property LO7 is unacceptable for certain kinds of motivational attitudes but a necessity for others. It is for instance perfectly possible for agents to have contradicting wishes[4], but it seems hardly rational to allow agents to try and fulfil these conflicting wishes simultaneously. Thus, whereas the absence of LO7 is essential when formalising wishes, the presence is when formalising goals.

It turns out that our formalisation of wishes validates all but one of the properties of logical omniscience.

6.8. Proposition. *All of the properties of logical omniscience formalised in Definition 6.7, with the exception of LO7, are valid for the* $\mathbf{W}_i$ *operator.*

Although we argued against the properties of logical omniscience when formalising motivational attitudes, we do not consider it a serious problem that our formalisation of wishes validates (almost all of) these properties. For these wishes are both *implicit* in the terminology of Levesque [80] and *passive* in the sense of Castelfranchi *et al.* [17]. Being implicit, it will not be the case that agents *explicitly* desire all of their wishes[5]. Being passive, wishes in themselves do not *actively* influence the course of action that an agent is going to take. Through the act of selecting, agents turn some of their implicit, passive wishes into explicit, active goals. Hence even though an agent implicitly and passively desires all logical consequences of one of its wishes, it will not do so explicitly

[3] The problem of unrestricted weakening is intuitively related to the Ross's paradox [115], well-known in deontic logic[3, 97]. The standard counterexample towards the desirability of LO6 in a deontic context, where the operator $\mathbf{X}$ is interpreted as 'being obliged to', is that of an agent that is obliged to mail a letter while not being obliged to either mail the letter or burn it.

[4] Even stronger, human agents will almost always suffer from conflicts between their wishes.

[5] For the implicit belief that, in combination with awareness, constitutes explicit belief in the approach of Fagin & Halpern [31], it is also considered unproblematic that the properties of logical omniscience are validated.

and actively. Therefore Proposition 6.8 is not taken to represent a severe problem for a formalisation of (implicit and passive) wishes, whereas it would for a formalisation of (explicit and active) goals. In the following section it will be shown how the properties of logical omniscience are avoided for goals.

6.3 Setting goals

As remarked previously, an agent's goals are not primitive but induced by its wishes. Basically, an agent selects among its (implicit and passive) wishes those that it (explicitly and actively) aims to fulfil. Given the rationality of agents, these selected wishes should be both unfulfilled and implementable: it does not make sense for an agent to try and fulfil a wish that either already has been fulfilled or for which fulfilment is not a practical possibility. We do not take the latter constraint too stringently, i.e. we only demand wishes to be individually implementable without requiring a simultaneous implementability of all chosen wishes. However, if desired, constraints like simultaneous implementability are easily formulated. The act of selecting is treated as a fully-fledged action by defining the opportunity, ability and result of selecting. Informally, an agent has the *opportunity* to select any of its wishes, corresponding to the idea that choices are only restricted by the elements among which is to be chosen. However, an agent is *capable* of selecting only those formulae that are unfulfilled and implementable, which can be thought of as it having a built-in aversion against selecting fulfilled or practically impossible formulae. The *result* of a selection will consist of the selected formula being marked chosen.

The notion of unfulfilledness is straightforwardly formalised as 'not holding', i.e. a formula φ is unfulfilled in a state s of some model M if and only if $\mathrm{M}, s \not\models^{\mathrm{C}} \varphi$. Defining implementability is a little more elaborate. Roughly speaking, we define a formula φ to be implementable for an agent i, denoted by $\Diamond_i\varphi$, if i has the practical possibility to fulfil φ by performing an appropriate sequence of atomic actions[6].

6.9. Definition. The binary relation $\models^{\mathrm{C}}$ between a formula in L^{C} and a pair M, s consisting of a model M for L^{C} and a state s in M is for implementability formulae defined by:

$$\mathrm{M}, s \models^{\mathrm{C}} \Diamond_i\varphi \Leftrightarrow \exists k \in \mathbb{N} \exists a_1, \ldots, a_k \in \mathrm{At}(\mathrm{M}, s \models^{\mathrm{C}} \mathbf{PracPoss}_i(a_1; \ldots; a_k, \varphi))$$

[6] As was pointed out by Maarten de Rijke, defining the implementability operator in this way makes it a kind of dual master modality (cf. [40, 122]). A formula consisting of a formula φ prefixed by the master modality is true in some state s of a model iff φ holds at all states that are reachable by any finite sequence of transitions from s. Such a formula is false iff there is some state s', reachable by some finite sequence of transitions from s, at which φ does not hold. This indeed makes our implementability modality to be a dual master modality.

Having defined unfulfilledness and implementability, we can now formally introduce the `select` action.

6.10. Definition. For $\mathrm{M} \in \mathbf{M}^{\mathrm{C}}$ with state s, $i \in \mathrm{A}$ and $\varphi \in \mathrm{L}$ we define:

$$\mathtt{r}^{\mathrm{C}}(i, \mathtt{select}\,\varphi)(\mathrm{M}, s) = \begin{cases} \emptyset & \text{if } \mathrm{M}, s \models^{\mathrm{C}} \neg \mathbf{W}_i\varphi \\ \mathtt{choose}(i, \varphi)(\mathrm{M}, s), s & \text{if } \mathrm{M}, s \models^{\mathrm{C}} \mathbf{W}_i\varphi \end{cases}$$

where for $\mathrm{M} = \langle \mathrm{S}, \pi, \mathrm{R}, \mathtt{r}_0, \mathtt{c}_0, \mathrm{W}, \mathrm{C}, \mathrm{Agenda} \rangle$ we define

$$\mathtt{choose}(i, \varphi)(\mathrm{M}, s) = \langle \mathrm{S}, \pi, \mathrm{R}, \mathtt{r}_0, \mathtt{c}_0, \mathrm{W}, \mathrm{C}', \mathrm{Agenda} \rangle \text{ with}$$
$$\mathrm{C}'(i', s') = \mathrm{C}(i', s') \text{ if } i \neq i' \text{ or } s \neq s'$$
$$\mathrm{C}'(i, s) \;\; = \mathrm{C}(i, s) \cup \{\varphi\}$$

$$\mathtt{c}^{\mathrm{C}}(i, \mathtt{select}\,\varphi)(\mathrm{M}, s) = \mathbf{1} \Leftrightarrow \mathrm{M}, s \models^{\mathrm{C}} \neg\varphi \wedge \Diamond_i\varphi$$

The binary relation $\models^{\mathrm{C}}$ between a formula in L^{C} and a pair M, s consisting of a model M for L^{C} and a state s in M is for choices defined by:

$$\mathrm{M}, s \models^{\mathrm{C}} \mathbf{C}_i\varphi \Leftrightarrow \varphi \in \mathrm{C}(i, s)$$

The definition of $\mathtt{r}^{\mathrm{C}}$ for the selection actions indeed provides for a correct model-transformation.

6.11. Proposition. *For all* $\mathrm{M} \in \mathbf{M}^{\mathrm{C}}$ *with state* s*, for all* $i \in \mathrm{A}$ *and* $\varphi \in \mathrm{L}$*, if* $\mathrm{M}', s = \mathtt{r}^{\mathrm{C}}(i, \mathtt{select}\,\varphi)(\mathrm{M}, s)$ *then* $\mathrm{M}' \in \mathrm{M}^{\mathrm{C}}_{\sim}$.

Besides being correct in that well-defined models are transformed into well-defined models, our formalisation of the act of selecting is also correct with respect to minimal change. That is, the change caused by selecting some formula is minimal given that the formula is to be marked chosen, which implies that our formalisation of selections does not suffer from the frame problem. The following proposition provides a (partial) formalisation of this property.

6.12. Proposition. *For all* $\mathrm{M} \in \mathbf{M}^{\mathrm{C}}$ *with state* s*, for all* $i \in \mathrm{A}$ *and* $\varphi \in \mathrm{L}$*, if* $\mathrm{M}', s = \mathtt{r}^{\mathrm{C}}(i, \mathtt{select}\,\varphi)(\mathrm{M}, s)$ *then for all states* s' *in* M*,* $\mathrm{M}, s' \models^{\mathrm{C}} \psi$ *iff* $\mathrm{M}', s' \models^{\mathrm{C}} \psi$*, for all* $\psi \in \mathrm{L}$.

Proposition 6.12 states that all formulae from L are interpreted identically in a model M and in the one resulting from selecting some formula in an arbitrary state of M. As a direct consequence of this proposition we have the following corollary, which states that the interpretation of wishes and implementability formulae persists under selecting some formula.

6.13. Corollary. *For all* $\mathrm{M} \in \mathbf{M}^{\mathrm{C}}$ *with state* s*, for all* $i \in \mathrm{A}$ *and* $\varphi \in \mathrm{L}$*, if* $\mathrm{M}', s = \mathrm{r}^{\mathrm{C}}(i, \mathtt{select}\,\varphi)(\mathrm{M}, s)$ *then for all states* s' *in* M *and all* $\psi \in \mathrm{L}$*:*

- $\mathrm{M}, s' \models^{\mathrm{C}} \mathbf{W}_i\psi \Leftrightarrow \mathrm{M}', s' \models^{\mathrm{C}} \mathbf{W}_i\psi$
- $\mathrm{M}, s' \models^{\mathrm{C}} \Diamond_i\psi \Leftrightarrow \mathrm{M}', s' \models^{\mathrm{C}} \Diamond_i\psi$

Having defined wishes and selections, one might be tempted to straightforwardly define goals to be selected wishes, i.e. $\mathbf{Goal}_i\varphi \triangleq \mathbf{W}_i\varphi \wedge \mathbf{C}_i\varphi$. This definition is however not adequate to formalise the idea of goals being selected, *unfulfilled*, *implementable* wishes. The reason for this is that in well-defined models from $\mathbf{M}^{\mathrm{C}}$ no relation is imposed between 'being selected' and 'being unfulfilled and implementable', i.e. one is not prevented by Definition 6.4 to come up with a well-defined model M in which for certain i and s the set $\mathrm{C}(i, s)$ contains formulae φ that are either fulfilled or not implementable. We see basically two ways of solving this problem, a semantical and a syntactical one. Semantically one could restrict the set of well-defined models for L^{C} to those in which the set $\mathrm{C}(i, s)$ contains for all agents i and states s only unfulfilled and implementable formulae, thereby ensuring beforehand that goals are unfulfilled and implementable when using the definition suggested above. Syntactically one could define goals to be only those selected wishes that are indeed unfulfilled and implementable. Hence instead of (semantically) restricting the set of well-defined models for L^{C} one (syntactically) expands the definition of goals. Although both the semantic and the syntactic approach are equally well applicable, we will restrict ourselves here to pursuing the syntactic one. Therefore, goals are defined to be those wishes that are unfulfilled, implementable and selected.

6.14. Definition. The $\mathbf{Goal}_i$ operator is for $i \in \mathrm{A}$ and $\varphi \in \mathrm{L}$ defined by:

$\mathbf{Goal}_i\varphi \triangleq \mathbf{W}_i\varphi \wedge \neg\varphi \wedge \Diamond_i\varphi \wedge \mathbf{C}_i\varphi$

As mentioned above, the goals of agents, being the explicit and active notions that they are, are not to validate the properties of logical omniscience as formalised in Definition 6.7. Fortunately, though not surprisingly, this indeed turns out to be the case when defining goals as in Definition 6.14.

6.15. Proposition. *None of the properties of logical omniscience formalised in Definition 6.7, with the exception of LO7, is valid for the* $\mathbf{Goal}_i$ *operator.*

The only property of logical omniscience satisfied by the goal operator, viz. LO7, formalises the idea that an agent's goals are consistent. This is a highly desirable property for rational creatures. For although it is quite possible for a rational agent to have contradictory wishes, it is rather irrational to try and fulfil these simultaneously.

Besides invalidating the undesired ones among the properties of logical omniscience, particularly those embodying the side-effect and transference problem, our definition of goals and selections has some other pleasant and desirable features. The following proposition formalises some of these features together with some properties characterising the act of selecting.

6.16. Proposition. *For all* $i \in \mathrm{A}$ *and* $\varphi \in \mathrm{L}$ *we have:*

1. $\models^{\mathrm{C}} \mathbf{W}_i\varphi \leftrightarrow \langle \mathrm{do}_i(\mathtt{select}\,\varphi)\rangle\top$
2. $\models^{\mathrm{C}} \langle \mathrm{do}_i(\mathtt{select}\,\varphi)\rangle\top \leftrightarrow \langle \mathrm{do}_i(\mathtt{select}\,\varphi)\rangle\mathbf{C}_i\varphi$
3. $\models^{\mathrm{C}} \neg\mathbf{A}_i\mathtt{select}\,\varphi \rightarrow [\mathrm{do}_i(\mathtt{select}\,\varphi)]\neg\mathbf{Goal}_i\varphi$
4. $\models^{\mathrm{C}} \mathbf{PracPoss}_i(\mathtt{select}\,\varphi, \top) \rightarrow \langle \mathrm{do}_i(\mathtt{select}\,\varphi)\rangle\mathbf{Goal}_i\varphi$
5. $\models^{\mathrm{C}} \varphi \Rightarrow \models^{\mathrm{C}} \neg\mathbf{Goal}_i\varphi$
6. $(\varphi \rightarrow \psi) \rightarrow (\mathbf{Goal}_i\varphi \rightarrow \mathbf{Goal}_i\psi)$ *is not for all* $\varphi, \psi \in \mathrm{L}$ *valid*
7. $\mathbf{K}_i(\varphi \rightarrow \psi) \rightarrow (\mathbf{Goal}_i\varphi \rightarrow \mathbf{Goal}_i\psi)$ *is not for all* $\varphi, \psi \in \mathrm{L}$ *valid*

The first item of Proposition 6.16 states that agents have the opportunity to select all, and nothing but, their wishes. The second item formalises the idea that every choice for which an agent has the opportunity results in the selected wish being marked chosen. In the third item it is stated that whenever an agent is unable to select some formula, then selecting this formula will not result in it becoming one of its goals. The related item 4 states that all, and nothing but, practically possible selections result in the chosen formula being a goal. The fifth item provides a strengthening of the invalidation of the second property of logical omniscience, which embodies the transference problem. It states that no logically inevitable formula qualifies as a goal. Hence whenever a formula is valid this does not only not necessarily imply that it is a goal but it even necessarily implies that it is not. The last two items of Proposition 6.16 are related to the avoidance of the transference problem, and state that goals are neither closed under implications nor under known implications.

6.4 Formalising commitments

The last part of our formalisation of motivational attitudes concerns the agents' commitments. Commitments to actions represent promises to perform these actions, i.e. an agent that is committed to an action has promised itself to perform the action. As mentioned above, commitments may be made to plans for goals, i.e. whenever an agent is committed it should be to an action that is correct and feasible to bring about at least one of its goals.

Not only do we formalise this static aspect of made commitments, but we also consider the dynamic aspect of making and undoing commitments. The act of committing is

related to, and can be seen as, an elementary implementation of practical reasoning, the process through which agents decide that they should perform certain actions (their ought-to-do's) on the basis of their wishes, desires or goals (their ought-to-be's). Ever since Aristotle, the study of practical reasoning has formed a major constituent of the research in analytical philosophy [111]. According to Von Wright [134], the essence of practical reasoning is best captured by the following syllogism:

> i intends to make it true that φ
>
> i thinks that, unless it does α, it will not achieve this
>
> Therefore i intends to do α.

The simplified version of practical reasoning that we aim to formalise through the act of committing can be described by the following syllogism,

> i *knows* that φ is one of its *goals*
>
> i *knows* that α is *correct* and *feasible* with respect to φ
>
> Therefore i has the *opportunity* to *commit* itself to α

which corresponds to the idea that commitments may be made to actions that are known to be correct and feasible to achieve some of the agent's goals.

Commitments are formalised through the $\mathbf{Committed}\_\_$ operator: $\mathbf{Committed}_i\alpha$ denotes that agent i is committed to the action α. The act of committing is modelled by the (special) action $\texttt{commit\_to}\_$: $\texttt{commit\_to}\,\alpha$ represents the act of committing to the (regular) action α. As mentioned in Section 6.1, commitments, though in general persistent, should not be maintained when having become useless or impossible, i.e. agents should have the possibility to undo useless or impossible commitments. This act of uncommitting is formalised by the $\texttt{uncommit}\_$ action: $\texttt{uncommit}\,\alpha$ denotes the act of undoing the commitment to the action α. In the sequel we successively formalise the act of committing, the commitments that have been made, and the act of uncommitting.

6.4.1 Getting committed

The act of committing, though of a special nature compared to other actions, is treated as a fully-fledged action, i.e. we define what it means to have the ability or opportunity to commit, and what the result of committing is. To start with the latter notion, given the relation between the infinitive 'to commit' and the past participle 'committed', it seems rather obvious that the act of committing should result in the agent being committed. Determining when an agent has the opportunity to perform a $\texttt{commit\_to}\,\alpha$ action is equally obvious, for it is inspired by the syllogism describing our version of

practical reasoning given above. Hence agent i has the opportunity to perform the action $\texttt{commit\_to}\,\alpha$ if and only if it knows that α is correct and feasible to bring about one of its goals. This leaves to determine the constituents of the ability of an agent to commit itself. Our definition of this ability is inspired by the observation that situations where agents are committed to two (or more) essentially different actions are highly problematic. Since 'being committed to α' intuitively corresponds to 'having promised (to oneself) to perform α next', it is unclear how to interpret the case where an agent is committed to two different actions. Should both actions be performed simultaneously? But what does it mean that actions are performed simultaneously? Are they performed concurrent, interleaved or in parallel? Or should the actions be performed sequentially? If so, in which order? And what then if the commitment to perform one action does not persist under execution of the other action? As an answer to these questions we propose that situations where agents have multiple commitments are to be avoided. One way to ensure this is to let an agent have the ability to commit itself only if it is not up to any previously made commitments, i.e. an agent is capable to commit only if it is not already committed.

As mentioned previously, an agent's commitments are interpreted by means of the so-called agenda function. The idea is that this function yields, for a given agent and a given state, the actions that the agent is committed to. Whenever an agent successfully commits itself to an action the agent's agenda is updated accordingly. The actual formal definition capturing this fairly unsophisticated idea is itself rather complicated. The reason for this lies in various desiderata that commitments and the act of committing should meet.

The first of these desiderata is that commitments should be known, i.e. agents should be aware of the commitments that they have made. To bring about this knowledge of commitments, epistemic equivalence classes rather than states are considered in an agenda update. Thus whenever agent i commits itself to action α in some state s of a model, the agenda of all states s' that are epistemically equivalent with s is updated appropriately.

The second and very important desideratum imposed on commitments is that they behave compositionally correct, i.e. the commitment to a composite action is linked in a rational way to commitments to its constituents. It is for example desirable that an agent that is committed to an action $\texttt{if}\,\varphi\,\texttt{then}\,\alpha_1\,\texttt{else}\,\alpha_2\,\texttt{fi}$ is also committed to α_1 whenever it knows that φ holds, and that an agent committed to the action $\alpha_1;\alpha_2$ is (currently) committed to α_1 and committed to α_2 in the state of affairs that results from executing α_1. To ensure the kind of rational behaviour associated with the conditional composition with known condition, an agent's agenda does not contain syntactical representations of made commitments, but instead the *semantic essence* of such a commitment. Consequently, the act of committing should not result in an update with the actual action that is

committed to, but instead adds the semantic essence of the newly made commitment to the agenda. This semantic essence, which is a situated notion dependent on an agent and a state, is given by a 'normalised' form of the (unique) finite computation run of the action for the agent in the state. As mentioned in Chapter 4, this finite computation run is a sequence of atomic actions and tests which constitutes the halting execution of an event in a state. A normal form of a basic action α is an action α' that originates from α by removing all brackets occurring in α and re-inserting them starting from the right. For example, a basic action $(a_1;a_2);a_3$ is normalised to $a_1;(a_2;a_3)$ and $(a_1;a_2);(a_3;a_4)$ is normalised to $a_1;(a_2;(a_3;a_4))$. By considering the (semantic) notion of normalised finite computation runs rather than the actions themselves, it is ensured that commitments depend on meaning rather than syntactical shape: if an agent is committed to an action it is also committed to all actions that are essentially identical. To bring about rational behaviour of commitments with respect to sequentially composed actions the actual update does not just concern the epistemic equivalence class of the current state, but also that of all the states that lay alongside the execution trajectory of the action. For example, if an agent i commits itself to $\alpha_1;\alpha_2$ in the state s of some model, then the epistemic equivalence class of s is updated with the commitment to α_1, and the epistemic equivalence class of the state s'' that results from executing α_1 in some s' that is an element of the epistemic equivalence of s is updated with the commitment to α_2.

Since the actions from Ac are deterministic, for each event built out of these actions there is at most one finite computation sequence which consists of the semi-atomic actions that occur in the halting executing of the event. Or phrased differently, the set of finite computation runs of a given event $\mathrm{do}_i(\alpha)$ is either empty or a singleton set. This property of deterministic actions facilitates the definition of finite computation runs to a considerable extent: simply define it to be the unique finite computation sequence for which execution terminates (compare this to the rather complex definition given in Chapter 4).

6.17. Definition. Since Ac is closed under the core clauses only, the function $\mathrm{CS} : \mathrm{Ac} \to \wp(\mathrm{Ac_b})$ is defined as usual. For $\mathrm{M} \in \mathbf{M}^{\mathrm{C}}$ the function $\mathrm{CR}^{\mathrm{C}}_{\mathrm{M}} : \mathrm{A} \times \mathrm{Ac} \times \mathrm{S} \to \wp(\mathrm{Ac_b})$ is defined by:

$$\mathrm{CR}^{\mathrm{C}}_{\mathrm{M}}(i,\alpha,s) = \{\alpha' \in \mathrm{CS}(\alpha) \mid \mathtt{r}^{\mathrm{C}}(i,\alpha')(\mathrm{M},s) \neq \emptyset\}$$

We will not bother to give a rigid mathematical definition of the $\mathrm{normalised}$ function, which turns basic actions into their normal form. The inherent simplicity of this function makes it not worthwhile to put an effort into defining it formally. We therefore assume the $\mathrm{normalised}$ function to be given, and do the same for the projection function π_2, which is assumed to yield the second element of a pair.

For reasons of convenience we introduce, analogously to the Can-predicate, a so-called Intend-predicate, which is meant to formalise the intentions of agents. The definition of

this predicate is based on the idea that agents (loosely) intend to do all the actions that are correct and feasible with respect to some of their goals. As such, intention provides the precondition for successful commitment[7].

6.18. Definition. For $\alpha \in \mathrm{Ac}^{\mathrm{C}}, i \in \mathrm{A}$ and $\varphi \in \mathrm{L}$ we define:

$\mathbf{Intend}_i(\alpha, \varphi) \triangleq \mathbf{Can}_i(\alpha, \varphi) \wedge \mathbf{K}_i\mathbf{Goal}_i\varphi$

Having established the formal prerequisites, we can now present the definitions formalising the intuitive description of the act of committing as presented above.

6.19. Definition. For all $\mathrm{M} \in \mathbf{M}^{\mathrm{C}}$ with state s, for all $i \in \mathrm{A}$ and $\alpha \in \mathrm{Ac}$ we define:

$\mathtt{r}^{\mathrm{C}}(i, \mathtt{commit\_to}\,\alpha)(\mathrm{M}, s) = \emptyset$ if $\mathrm{M}, s \models^{\mathrm{C}} \neg\mathbf{Intend}_i(\alpha, \varphi)$ for all $\varphi \in \mathrm{C}(i, s)$
$\mathtt{r}^{\mathrm{C}}(i, \mathtt{commit\_to}\,\alpha)(\mathrm{M}, s) = \mathrm{M}', s$ with $\mathrm{M}' = \langle \mathrm{S}, \pi, \mathrm{R}, \mathtt{r}_0, \mathtt{c}_0, \mathrm{W}, \mathrm{C}, \mathrm{Agenda}' \rangle$
where for all $s' \in [s]_{\mathrm{R}(i)}$, $\mathrm{Agenda}'(i, s') = \mathrm{Agenda}(i, s') \cup \mathrm{normalised}(\mathrm{CR}^{\mathrm{C}}_{\mathrm{M}}(i, \alpha, s'))$
and for all $1 \leq k \leq m \Leftrightarrow 1$, $s' \in [s]_{\mathrm{R}(i)}$
$\mathrm{Agenda}'(i, s'') = \mathrm{Agenda}(i, s'') \cup \{\beta_{k+1}; \ldots; \beta_m\}$ where
$s'' \in [\pi_2(\mathtt{r}^{\mathrm{C}}(i, \beta_1; \ldots \beta_k)(\mathrm{M}, s'))]_{\mathrm{R}(i)}$ for $\beta_1; \ldots; \beta_m = \mathrm{normalised}(\mathrm{CR}^{\mathrm{C}}_{\mathrm{M}}(i, \alpha, s'))$
otherwise

$\mathtt{c}^{\mathrm{C}}(i, \mathtt{commit\_to}\,\alpha)(\mathrm{M}, s) = \mathbf{1}$ iff $\mathrm{Agenda}(i, s) = \emptyset$

To make Definition 6.19 come to life, and in particular to shed some light on the rather abstract and fairly complicated definition of $\mathtt{r}^{\mathrm{C}}$ as it is given above, consider Figure 6.1.

Figure 6.1 is a pictorial representation of what happens when an agent i makes a successful commitment to an action α in the leftmost state s of the model represented in the figure. The small circles represent different states of the model, the leftmost arcs, annotated with $\mathrm{R}(i)$, represent elements of i's epistemic accessibility relation, the dotted squares denote epistemic equivalence classes of i centred around a state, and the arcs annotated with α_j, β_j or γ_j represent transitions between states. We assume that the action α, to which i commits itself in s, is such that

- $\mathrm{normalised}(\mathrm{CR}^{\mathrm{C}}_{\mathrm{M}}(i, \alpha, s_0)) = \alpha_1; (\alpha_2; (\ldots (\alpha_{k-1}; \alpha_k) \ldots))$
- $\mathrm{normalised}(\mathrm{CR}^{\mathrm{C}}_{\mathrm{M}}(i, \alpha, s)) = \beta_1; (\beta_2; (\ldots (\beta_{l-1}; \beta_l) \ldots))$
- $\mathrm{normalised}(\mathrm{CR}^{\mathrm{C}}_{\mathrm{M}}(i, \alpha, s_1)) = \gamma_1; (\gamma_2; (\ldots (\gamma_{m-1}; \gamma_m) \ldots))$

In accordance with Definition 6.19, i's commitment to α is carried out by updating the agenda of i in each state that is an element of an epistemic equivalence class appearing alongside the execution trajectory of α. In terms of Figure 6.1 this comes down to extending the agenda of each state in A_j with $\alpha_{j+1}; (\ldots (\alpha_{k-1}; \alpha_k) \ldots)$, while causing

[7] Our paraphrase of Cohen & Levesque's motto 'intention is choice plus commitment' [20] could therefore be stated as 'commitments are chosen intentions'.

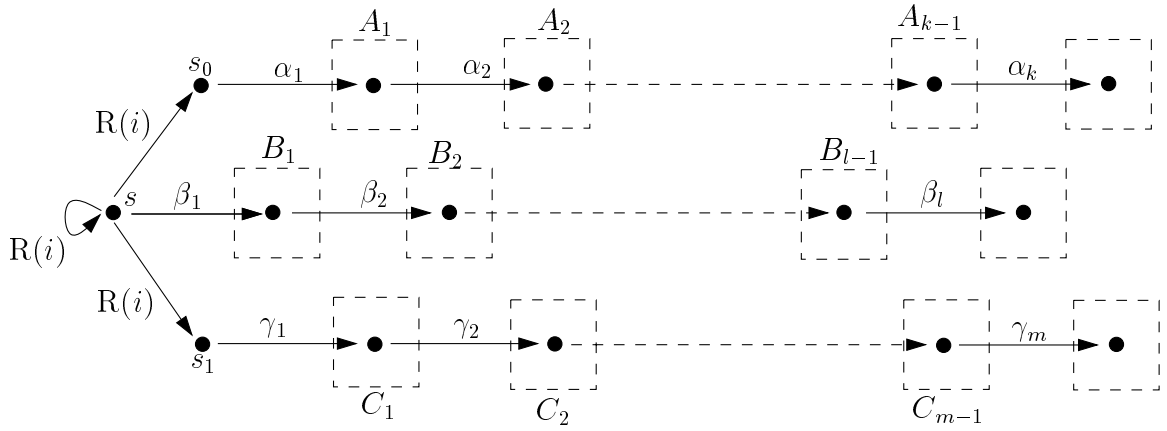

FIGURE 6.1. The act of committing

analogous changes to the B_j and C_j equivalence classes. Some aspects of Figure 6.1 are particular worth noticing, the first of these being the fact that, except for the epistemic equivalence class surrounding the initial states, all agenda updates are carried out on the level of epistemic equivalence classes rather than on the level of states. That is, the agenda of i is extended with one and the same action in all the states of such an epistemic equivalence class. Moreover, the actual nature of this action is determined and fixed in the equivalence class of the initial state. A last but important point to notice is that only the agenda of i is modified, and that only in those states that are somehow, i.e. by a combination of state-transitions and epistemic accessibility relations, connected to the state in which the commitment is being made. All other elements of the model remain unchanged.

The latter aspect mentioned above, i.e. the minimality of the change caused by performing a commitment, is partly formalised in Proposition 6.21 given below. Proposition 6.20 states the correctness of the definition of r^{C} as presented above in the sense that it yields a (unique) well-defined model when applied to a well-defined model.

6.20. PROPOSITION. *For all* $\mathrm{M} \in \mathbf{M}^{\mathrm{C}}$ *with state* s*, for all* $i \in \mathrm{A}$ *and* $\alpha \in \mathrm{Ac}$*, if* $\mathrm{M}', s = \mathrm{r}^{\mathrm{C}}(i, \mathtt{commit\_to}\,\alpha)(\mathrm{M}, s)$ *then* $\mathrm{M}' \in \mathrm{M}^{\mathrm{C}}_{\sim}$.

6.21. PROPOSITION. *For all* $\mathrm{M} \in \mathbf{M}^{\mathrm{C}}$ *with state* s*, for all* $i \in \mathrm{A}$ *and* $\alpha \in \mathrm{Ac}$*, if* $\mathrm{M}', s = \mathrm{r}^{\mathrm{C}}(i, \mathtt{commit\_to}\,\alpha)(\mathrm{M}, s)$ *then for all states* s' *in* M, $\mathrm{M}, s' \models^{\mathrm{C}} \varphi$ *iff* $\mathrm{M}', s' \models^{\mathrm{C}} \varphi$*, for all* $\varphi \in \mathrm{L}$.

Additional properties related to the commit actions are given in 6.4.4.

6.4.2 Being committed

After the rather elaborate and fairly complicated definition formalising the act of committing, defining what it means to be committed is a relatively straightforward and easy job. Basically, agents are committed to all actions whose semantic essence is captured by the normalised basic action in the appropriate agenda. The only additional aspect that has to be taken into account when defining the semantics of the $\mathbf{Committed\_}$ operator is that agents should start at the very beginning (a very good place to start), i.e. whenever an agent's agenda contains a normalised basic action which is not semi-atomic, then the agent is also committed to actions whose semantic essence is a prefix of this normalised basic action. This constraint is quite an obvious one: how can agents be faithfully committed to a sequentially composed action if not committed to its first constituent? Formally we ensure this behaviour by using the prefix relation on basic actions. The definition of $\models^{\mathrm{C}}$ for the $\mathbf{Committed}_i$ operator could then be informally interpreted as 'an agent is committed to those actions of which the semantic essence is a prefix of one of the actions in its agenda'.

6.22. Definition. The binary relation $\models^{\mathrm{C}}$ between a formula in L^{C} and a pair M, s consisting of a model M for L^{C} and a state s in M is for commitments defined by:

$$\mathrm{M}, s \models^{\mathrm{C}} \mathbf{Committed}_i\alpha \Leftrightarrow$$
$$\forall s' \in [s]_{\mathrm{R}(i)} \exists \alpha_1 \in \mathrm{CR}^{\mathrm{C}}_{\mathrm{M}}(i, \alpha, s') \exists \alpha_2 \in \mathrm{Agenda}(i, s')(\mathrm{Prefix}(\mathrm{normalised}(\alpha_1), \alpha_2))$$

An investigation of the properties of the commitment operator is postponed to 6.4.4.

6.4.3 Getting uncommitted

By performing an uncommit action, agents may undo previously made commitments that turned out to be either useless or impossible. That is, as soon as an agent no longer knows some commitment to be correct and feasible for at least one of its goals it may undo this commitment. Just as we did for the commit action, we have to decide upon the constituents of the result, opportunity and ability for the actions formalising the act of uncommitting. The result of such an action is obvious: agents should no longer be committed to α after a successful performance of an $\mathtt{uncommit}\,\alpha$ action[8]. Defining what it means to have the opportunity and ability to uncommit represents a somewhat more arbitrarily choice. We have decided to let an agent have the opportunity to undo

[8] As was pointed out to me by John Fox, this description of the result of undoing a commitment comprises a major simplification. For in real life, undoing commitments involves more than just abandoning future commitments: it is also necessary to (try to) undo all the effects that followed from initially pursuing the commitment. For example, if an agent that is committed to $\alpha_1; \alpha_2$ finds out after having done α_1 that its commitment to α_2 should be undone, then it should not only remove α_2 from its agenda but also try to undo as many of the effects of α_1 as possible.

any of its commitments, i.e. there is nothing in its circumstances that may prevent an agent to undo a commitment. Our loyal, diligent agents are however only (morally) capable of undoing commitments that have become redundant. The actual definition of the functions $\mathtt{r}^{\mathrm{C}}$ and $\mathtt{c}^{\mathrm{C}}$ consists of nothing but a formalisation of these intuitive ideas.

6.23. Definition. For all $\mathrm{M} \in \mathbf{M}^{\mathrm{C}}$ with state s, for all $i \in \mathrm{A}$ and $\alpha \in \mathrm{Ac}$ we define:

$\mathtt{r}^{\mathrm{C}}(i, \mathtt{uncommit}\,\alpha)(\mathrm{M}, s) = \emptyset$ if $\mathrm{M}, s \models^{\mathrm{C}} \neg\mathbf{Committed}_i\alpha$
$\mathtt{r}^{\mathrm{C}}(i, \mathtt{uncommit}\,\alpha)(\mathrm{M}, s) = \mathrm{M}', s$ with $\mathrm{M}' = \langle \mathrm{S}, \pi, \mathrm{R}, \mathtt{r}_0, \mathtt{c}_0, \mathrm{W}, \mathrm{C}, \mathrm{Agenda}' \rangle$
  where for all $s' \in [s]_{\mathrm{R}(i)}$, $\mathrm{Agenda}'(i, s') = \mathrm{Agenda}(i, s') \setminus \mathrm{normalised}(\mathrm{CR}^{\mathrm{C}}_{\mathrm{M}}(i, \alpha, s'))$
  and for all $1 \leq k \leq m \Leftrightarrow 1$, $s' \in [s]_{\mathrm{R}(i)}$
    $\mathrm{Agenda}'(i, s'') = \mathrm{Agenda}(i, s'') \setminus \{\beta_{k+1}; \ldots; \beta_m\}$ where
      $s'' \in [\pi_2(\mathtt{r}^{\mathrm{C}}(i, \beta_1; \ldots \beta_k)(\mathrm{M}, s'))]_{\mathrm{R}(i)}$ for $\beta_1; \ldots; \beta_m = \mathrm{normalised}(\mathrm{CR}^{\mathrm{C}}_{\mathrm{M}}(i, \alpha, s'))$
otherwise

$\mathtt{c}^{\mathrm{C}}(i, \mathtt{uncommit}\,\alpha)(\mathrm{M}, s) = \mathbf{1}$ iff $\mathrm{M}, s \models^{\mathrm{C}} \neg\mathbf{Intend}_i(\alpha, \varphi)$ for all $\varphi \in \mathrm{C}(i, s)$

Our definition of $\mathtt{r}^{\mathrm{C}}$ for the `uncommit` actions is also twofold correct: not only does performing an `uncommit` action provide for a correct model-transformation, but also does it do so while causing minimal change.

6.24. Proposition. *For all* $\mathrm{M} \in \mathbf{M}^{\mathrm{C}}$ *with state* s*, for all* $i \in \mathrm{A}$ *and* $\alpha \in \mathrm{Ac}$*, if* $\mathrm{M}', s = \mathtt{r}^{\mathrm{C}}(i, \mathtt{uncommit}\,\alpha)(\mathrm{M}, s)$ *then* $\mathrm{M}' \in \mathrm{M}^{\mathrm{C}}_{\sim}$.

6.25. Proposition. *For all* $\mathrm{M} \in \mathbf{M}^{\mathrm{C}}$ *with state* s*, for all* $i \in \mathrm{A}$ *and* $\alpha \in \mathrm{Ac}$*, if* $\mathrm{M}', s = \mathtt{r}^{\mathrm{C}}(i, \mathtt{uncommit}\,\alpha)(\mathrm{M}, s)$ *then for all states* s' *in* M, $\mathrm{M}, s' \models^{\mathrm{C}} \varphi$ *iff* $\mathrm{M}', s' \models^{\mathrm{C}} \varphi$*, for all* $\varphi \in \mathrm{L}$.

Additional validities characterising the `uncommit` action are given below.

6.4.4 The statics and dynamics of commitments

Here we characterise the statics and dynamics of commitments by presenting some validities for $\models^{\mathrm{C}}$. For a start we consider a number of validities characterising the dynamics of commitments.

6.26. Proposition. *For all* $i \in \mathrm{A}$, $\alpha, \beta \in \mathrm{Ac}$ *and* $\varphi \in \mathrm{L}$ *we have:*

1. $\models^{\mathrm{C}} \mathbf{Intend}_i(\alpha, \varphi) \rightarrow \langle \mathrm{do}_i(\mathtt{commit\_to}\,\alpha) \rangle \top$
2. $\models^{\mathrm{C}} \langle \mathrm{do}_i(\mathtt{commit\_to}\,\alpha) \rangle \top \leftrightarrow \langle \mathrm{do}_i(\mathtt{commit\_to}\,\alpha) \rangle \mathbf{Committed}_i\alpha$
3. $\models^{\mathrm{C}} \mathbf{Committed}_i\alpha \rightarrow \neg\mathbf{A}_i\mathtt{commit\_to}\,\beta$
4. $\models^{\mathrm{C}} [\mathrm{do}_i(\mathtt{commit\_to}\,\alpha)] \neg\mathbf{A}_i\mathtt{commit\_to}\,\beta$

5. $\models^{\mathrm{C}} \mathbf{Committed}_i\alpha \leftrightarrow \langle \mathrm{do}_i(\mathtt{uncommit}\,\alpha)\rangle\neg\mathbf{Committed}_i\alpha$
6. $\models^{\mathrm{C}} \mathbf{Intend}_i(\alpha,\varphi) \rightarrow \neg\mathbf{A}_i\mathtt{uncommit}\,\alpha$
7. $\models^{\mathrm{C}} \mathbf{A}_i\mathtt{uncommit}\,\alpha \leftrightarrow \mathbf{K}_i\mathbf{A}_i\mathtt{uncommit}\,\alpha$
8. $\models^{\mathrm{C}} \mathbf{Committed}_i\alpha \wedge \neg\mathbf{Can}_i(\alpha,\top) \rightarrow \mathbf{Can}_i(\mathtt{uncommit}\,\alpha, \neg\mathbf{Committed}_i\alpha)$

The first two items of Proposition 6.26 jointly formalise our version of the syllogism of practical reasoning as described above. In the third item it is stated that being committed prevents an agent from having the ability to (re)commit. The fourth item states that the act of committing is ability-destructive with respect to future `commit` actions, i.e. by performing a commitment an agent loses its ability to make any other commitments. Item 5 states that being committed is a necessary and sufficient condition for having the opportunity to uncommit; as mentioned above, agents have the opportunity to undo all of their commitments. In item 6 it is stated that agents are (morally) unable to undo commitments to actions that are still known to be correct and feasible to achieve some goal. In item 7 it is formalised that agents know of their abilities to uncommit to some action. The last item states that whenever an agent is committed to an action that is no longer known to be practically possible, it knows that it can undo this impossible commitment.

The following proposition formalises some of the desiderata for the statics of commitments that turn out to be valid in the class $\mathbf{M}^{\mathrm{C}}$ of models for L^{C}.

6.27. Proposition. *For all $i \in \mathrm{A}$, $\alpha,\alpha_1,\alpha_2 \in \mathrm{Ac}$ and all $\varphi \in \mathrm{L}$ we have:*

1. $\models^{\mathrm{C}} \mathbf{Committed}_i\alpha \rightarrow \mathbf{K}_i\mathbf{Committed}_i\alpha$
2. $\models^{\mathrm{C}} \mathbf{Committed}_i\mathtt{if}\,\varphi\,\mathtt{then}\,\alpha_1\,\mathtt{else}\,\alpha_2\,\mathtt{fi} \wedge \mathbf{K}_i\varphi \rightarrow \mathbf{Committed}_i(\mathtt{confirm}\,\varphi;\alpha_1)$
3. $\models^{\mathrm{C}} \mathbf{Committed}_i\mathtt{if}\,\varphi\,\mathtt{then}\,\alpha_1\,\mathtt{else}\,\alpha_2\,\mathtt{fi} \wedge \mathbf{K}_i\neg\varphi \rightarrow \mathbf{Committed}_i(\mathtt{confirm}\,\neg\varphi;\alpha_2)$
4. $\models^{\mathrm{C}}\mathbf{Committed}_i\mathtt{while}\,\varphi\,\mathtt{do}\,\alpha\,\mathtt{od} \wedge \mathbf{K}_i\varphi \rightarrow$
   $\mathbf{Committed}_i((\mathtt{confirm}\,\varphi;\alpha);\mathtt{while}\,\varphi\,\mathtt{do}\,\alpha\,\mathtt{od})$

The first item of Proposition 6.27 states that commitments are known, and the second and third item formalise the rationality of agents with regard to their commitments to conditionally composed actions. The last item concerns the unfolding of a while-loop: if an agent is committed to a while-loop while knowing the condition of the loop to be true, then the agent is also committed to the then-part of the while-loop.

The attentive reader may have noticed the absence of any validities characterising the commitments to sequentially composed actions. However, recall that among the desiderata we formulated for commitments was one stating that whenever an agent is committed to an action $\alpha_1;\alpha_2$ it is also committed to α_1 (now) and to α_2 in the state of affairs following execution of α_1. This desideratum is for the greater part met by our formalisation. That is, it is indeed the case that an agent committed to $\alpha_1;\alpha_2$ is also committed to

α_1. A careful examination of Figure 6.1 reveals that it is not necessarily the case that a commitment to $\alpha_1;\alpha_2$ implies a commitment to α_2 in the state of affairs resulting from the execution of α_1. Consider for example some state s' in B_1 which differs from the one represented by the circle. In s' it holds that i is committed to $\beta_2;(\beta_3;(\ldots(\beta_{l-1};\beta_l)\ldots))$, i.e. $\mathrm{M},s' \models^{\mathrm{C}} \mathbf{Committed}_i(\beta_2;(\beta_3;(\ldots(\beta_{l-1};\beta_l)\ldots)))$. It is however by no means guaranteed that i is committed to $\beta_3;(\ldots(\beta_{l-1};\beta_l)\ldots)$ after performing β_2, i.e. M,s' does not necessarily satisfy $[\mathrm{do}_i(\beta_2)]\mathbf{Committed}_i(\beta_3;(\ldots(\beta_{l-1};\beta_l)\ldots))$. Despite the fact that we did not completely meet the demand formulated for commitments to sequentially composed actions, it turns out that we do so for the action central to a commitment, i.e. the action that the commitment was originally made to (in Figure 6.1 α is central to the commitment). The second item of Proposition 6.28 formalises this property.

6.28. Proposition. *For all* $i \in \mathrm{A}$ *and* $\alpha_1,\alpha_2 \in \mathrm{Ac}$ *we have:*

- $\models^{\mathrm{C}} \mathbf{Committed}_i(\alpha_1;\alpha_2) \rightarrow \mathbf{Committed}_i\alpha_1$
- $\models^{\mathrm{C}}\langle \mathrm{do}_i(\mathtt{commit\_to}\,(\alpha_1;\alpha_2))\rangle\top \rightarrow$
  $\langle \mathrm{do}_i(\mathtt{commit\_to}\,(\alpha_1;\alpha_2))\rangle\mathbf{K}_i[\mathrm{do}_i(\alpha_1)]\mathbf{Committed}_i\alpha_2$

The first item of Proposition 6.28 states that a commitment to $\alpha_1;\alpha_2$ implies one to α_1. The second item states that as the result of a successful commitment to $\alpha_1;\alpha_2$ an agent *knows* that it is committed to α_2 after performing α_1. As such this second item is a weaker variant of the non-validated part of the desideratum for commitments to sequentially composed actions.

6.5 Summary and conclusions

In this chapter we presented a formalisation of motivational attitudes, the attitudes that explain why agents act the way they do. This formalisation concerns operators both on the assertion level, where operators range over propositions, and on the practition level, where operators range over actions. An important feature of our formalisation is the attention paid to the acts associated with selecting between wishes and with (un)committing to actions. Starting from the primitive notion of wishes, we defined goals to be selected, unfulfilled, implementable wishes. Commitments may be made to actions that are known to be correct and feasible with respect to some goal and may be undone whenever the action to which an agent has committed itself has either become impossible or useless. Both the act of making, and the act of undoing commitments are formalised as model-transforming actions in our framework. The actions that an agent is committed to are recorded in its agenda in such a way that commitments are closed under prefix-taking and under practical identity. On the whole our formalisation is a rather expressive one, which tries to be faithful to a certain extent to both commonsense intuition and philosophical insights.

6.5.1 Possible extensions

The major extension to the framework presented in this chapter concerns a formalisation of the actual execution of actions. Although the conditional nature of a framework based on dynamic logic makes it perhaps less suitable for an adequate formalisation of 'doing', one could think of a practition operator indicating which action is actually performed next. Using this predicate would enhance expressiveness in that it would be possible to formulate relations between actions that agents are committed to, and actions that they actually perform. Another way to extend the framework would be by establishing further relations with deontic notions like obligations and violations. A combination of the 'doing'-predicate with a deontic notion modelling violations or penalties would then allow one to model that agents should execute the actions that they are committed to if they want to avoid penalties. Research along these lines was initiated by Dignum & Van Linder [25, 26].

6.5.2 Bibliographical notes

This chapter is a thoroughly revised version of [86], to which a more elaborate dynamic component is added and from which some questionable restrictions have been removed.

The formalisation of motivational attitudes has received much attention within the agent research community. Probably the most influential account of motivational attitudes is due to Cohen & Levesque [20]. Starting from the primitive notions of implicit goals and beliefs, Cohen & Levesque define so-called persistent goals, which are goals which agents give up only when they think they are either satisfied or will never be true, and intentions, both ranging over propositions and over actions. The idea underlying persistent goals is similar to that underlying our notion of goals. Agents intend to bring about a proposition if they intend to do some action that brings about the proposition. An agent intends to do an action if it has the persistent goal to have done the action. This reduction of intentions to do actions for goals is a rather artificial and philosophically very questionable one: although intentions to actions should be related to goals, this relation should express that doing the action helps in bringing about some goal and not that doing the action in itself is a goal. Furthermore the coexistence of goals and intentions ranging over propositions seems to complicate matters unnecessarily.

Another important formalisation of motivational attitudes is proposed by Rao & Georgeff [109] in their BDI-architecture. Treating desires and intentions as primitive, Rao & Georgeff focus on the process of intention revision rather than the 'commitment acquisition' which is essential to our formalisation. Both desires and intentions in their framework suffer from the problems associated with logical omniscience. To avoid these problems, Cavedon *et al.* [18] propose the use of non-normal logics of intention and belief

in the BDI-architecture, and more in particular Rantala's 'impossible worlds' framework [106]. This 'impossible worlds' approach was originally proposed as a way to solve the problems of logical omniscience for informational attitudes. Hence, whereas we more or less employ the awareness approach, Cavedon *et al.* propose yet another technique developed to solve the problems of logical omniscience. It therefore may come as no surprise that the properties that Cavedon *et al.* acquire for intentions are highly similar to the properties of goals given in Section 6.3.

The last formalisation of motivational attitudes that we would like to mention is the one proposed by Dignum *et al.* [27]. In this formalisation, which is inspired by and based on research on deontic logic as carried out by Dignum *et al.*, notions like decisions, intentions and commitments are modelled. Of these, decisions and the act of committing are interpreted as so-called meta-actions, a notion similar to that of model-transformers. Despite its complexity, which is due to the incorporation of an algebraic semantics of actions and a trace semantics to model histories, some of the essential ideas underlying the formalisation of Dignum *et al.* are not unlike those underlying the formalisation presented in this chapter.

6.6 Selected proofs

6.8. Proposition. *All of the properties of logical omniscience formalised in Definition 6.7, with the exception of LO7, are valid for the* $\mathbf{W}_i$ *operator.*

Proof: Properties LO1 and LO2 state that $\mathbf{W}_i$ is a normal modal operator and are shown as for any necessity operator. Property LO3 follows directly by combining LO1 and LO2, and LO4 is a direct consequence of LO3. Properties LO5 and LO6 are typical for necessity operators: for whenever both φ and ψ hold at a set of designated worlds, $\varphi \wedge \psi$ also holds at all the worlds from that set (LO5), and if φ holds at all worlds from some set then $\varphi \vee \psi$ does also (LO6). That LO7 is not valid for the $\mathbf{W}_i$ operator is seen by considering a model M with state s such that no state s' exists with $(s, s') \in \mathrm{W}(i)$. Then it holds that $\mathrm{M}, s \models^{\mathrm{C}} \mathbf{W}_i\varphi \wedge \mathbf{W}_i\neg\varphi$, for all $\varphi \in \mathrm{L}$.
⊠

6.12. Proposition. *For all* $\mathrm{M} \in \mathbf{M}^{\mathrm{C}}$ *with state* s*, for all* $i \in \mathrm{A}$ *and* $\varphi \in \mathrm{L}$*, if* $\mathrm{M}', s = \mathrm{r}^{\mathrm{C}}(i, \mathtt{select}\,\varphi)(\mathrm{M}, s)$ *then for all states* s' *in* M*,* $\mathrm{M}, s' \models^{\mathrm{C}} \psi$ *iff* $\mathrm{M}', s' \models^{\mathrm{C}} \psi$*, for all* $\psi \in \mathrm{L}$.

Proof: The 'official' proof of this proposition proceeds by an induction not unlike the one used in proving completeness of the logic LCap in Chapter 3. Unofficially, this proposition is obvious since the models M and M' as mentioned above agree completely

on all their elements that are used to interpret formulae from L. The only element in which they (possibly) differ, i.e. the function C, is not used in the interpretation of formulae from L.
⊠

6.15. Proposition. *None of the properties of logical omniscience formalised in Definition 6.7, with the exception of LO7, is valid for the* $\mathbf{Goal}_i$ *operator.*

Proof: Properties LO1, LO3, LO4, LO5 and LO6 are most easily seen not to hold for the goal operator by noting the absence of any closure properties on the set $\mathrm{C}(i,s)$, for $i \in \mathrm{A}$ and s some state. Due to this absence it is perfectly possible that φ and $\varphi \to \psi$ are both in $\mathrm{C}(i,s)$ while ψ is not (LO1), that $\varphi \in \mathrm{C}(i,s)$ and $\psi \notin \mathrm{C}(i,s)$ while $\models^{\mathrm{C}} \varphi \to \psi$ (LO3) or $\models^{\mathrm{C}} \varphi \leftrightarrow \psi$ (LO4), that $\{\varphi, \psi\} \subseteq \mathrm{C}(i,s)$ and $\varphi \wedge \psi \notin \mathrm{C}(i,s)$ (LO5), or that $\varphi \in \mathrm{C}(i,s)$ while $\varphi \vee \psi \notin \mathrm{C}(i,s)$ (LO6), for appropriate $i \in \mathrm{A}$ and s a state in some model. Property LO2 is seen not to hold by observing that $\models^{\mathrm{C}} \varphi$ implies that φ is fulfilled always and everywhere, which means that φ is not a goal. In fact, one can show that whenever φ is inevitable, i.e. $\models^{\mathrm{C}} \varphi$ holds, it is necessarily not a goal, i.e. $\models^{\mathrm{C}} \neg\mathbf{Goal}_i\varphi$ holds (cf. item 5 of Proposition 6.16).

That LO7 holds for goals is a direct consequence of their unfulfilledness. For in any possible state s of any possible model M, either φ holds and thereby $\mathrm{M}, s \not\models^{\mathrm{C}} \mathbf{Goal}_i\varphi$, or $\neg\varphi$ holds and thereby $\mathrm{M}, s \not\models^{\mathrm{C}} \mathbf{Goal}_i\neg\varphi$. Hence LO7 is a valid property for goals.
⊠

6.16. Proposition. *For all* $i \in \mathrm{A}$ *and* $\varphi \in \mathrm{L}$ *we have:*

1. $\models^{\mathrm{C}} \mathbf{W}_i\varphi \leftrightarrow \langle \mathrm{do}_i(\mathtt{select}\,\varphi)\rangle\top$
2. $\models^{\mathrm{C}} \langle \mathrm{do}_i(\mathtt{select}\,\varphi)\rangle\top \leftrightarrow \langle \mathrm{do}_i(\mathtt{select}\,\varphi)\rangle\mathbf{C}_i\varphi$
3. $\models^{\mathrm{C}} \neg\mathbf{A}_i\mathtt{select}\,\varphi \to [\mathrm{do}_i(\mathtt{select}\,\varphi)]\neg\mathbf{Goal}_i\varphi$
4. $\models^{\mathrm{C}} \mathbf{PracPoss}_i(\mathtt{select}\,\varphi, \top) \to \langle \mathrm{do}_i(\mathtt{select}\,\varphi)\rangle\mathbf{Goal}_i\varphi$
5. $\models^{\mathrm{C}} \varphi \Rightarrow \not\models^{\mathrm{C}} \neg\mathbf{Goal}_i\varphi$
6. $(\varphi \to \psi) \to (\mathbf{Goal}_i\varphi \to \mathbf{Goal}_i\psi)$ *is not for all* $\varphi, \psi \in \mathrm{L}$ *valid*
7. $\mathbf{K}_i(\varphi \to \psi) \to (\mathbf{Goal}_i\varphi \to \mathbf{Goal}_i\psi)$ *is not for all* $\varphi, \psi \in \mathrm{L}$ *valid*

Proof: We successively show all items. Let $\mathrm{M} \in \mathbf{M}^{\mathrm{C}}$ with state s and $\varphi \in \mathrm{L}$ be arbitrary.

1. An easy inspection of Definition 6.10 shows that $\mathtt{r}^{\mathrm{C}}(i, \mathtt{select}\,\varphi)(\mathrm{M}, s) = \emptyset$ iff $\mathrm{M}, s \not\models^{\mathrm{C}} \mathbf{W}_i\varphi$. Thus $\mathrm{M}, s \models^{\mathrm{C}} \mathbf{W}_i\varphi \leftrightarrow \langle \mathrm{do}_i(\mathtt{select}\,\varphi)\rangle\top$, which was to be shown.
2. If $\mathrm{M}', s = \mathtt{r}^{\mathrm{C}}(i, \mathtt{select}\,\varphi)(\mathrm{M}, s)$, then M' is such that $\mathrm{C}'(i,s)$ contains φ. Then by definition $\mathrm{M}', s \models^{\mathrm{C}} \mathbf{C}_i\varphi$, and thus $\mathrm{M}, s \models^{\mathrm{C}} \langle \mathrm{do}_i(\mathtt{select}\,\varphi)\rangle\mathbf{C}_i\varphi$ if $\mathrm{M}, s \models^{\mathrm{C}} \langle \mathrm{do}_i(\mathtt{select}\,\varphi)\rangle\top$, which suffices to conclude item 2.

3. Suppose $\mathrm{M}, s \models^{\mathrm{C}} \neg\mathbf{A}_i\mathtt{select}\,\varphi$, i.e. $\mathrm{M}, s \models^{\mathrm{C}} \varphi \vee \neg\Diamond_i\varphi$. Now by definition, $\varphi \in \mathrm{L}$, and hence, by Proposition 6.12, $\mathrm{M}', s \models^{\mathrm{C}} \varphi$ if $\mathrm{M}, s \models^{\mathrm{C}} \varphi$ whenever $\mathrm{M}', s = \mathtt{r}^{\mathrm{C}}(i, \mathtt{select}\,\varphi)(\mathrm{M}, s)$. By Corollary 6.13 it follows that for M' as aforementioned holds that $\mathrm{M}', s \models^{\mathrm{C}} \neg\Diamond_i\varphi$ if $\mathrm{M}, s \models^{\mathrm{C}} \neg\Diamond_i\varphi$. Thus if $\mathrm{M}, s \models^{\mathrm{C}} \varphi \vee \neg\Diamond_i\varphi$ then it holds for $\mathrm{M}', s = \mathtt{r}^{\mathrm{C}}(i, \mathtt{select}\,\varphi)(\mathrm{M}, s)$ that $\mathrm{M}', s \models^{\mathrm{C}} \varphi \vee \neg\Diamond_i\varphi$. By definition it then directly follows that $\mathrm{M}', s \models^{\mathrm{C}} \neg\mathbf{Goal}_i\varphi$, and thus $\mathrm{M}, s \models^{\mathrm{C}} \neg\mathbf{A}_i\mathtt{select}\,\varphi \rightarrow [\mathrm{do}_i(\mathtt{select}\,\varphi)]\neg\mathbf{Goal}_i\varphi$, which was to be shown.
4. This item follows by combining item 2 of this proposition with Proposition 6.12 and Corollary 6.13.
5. If $\models^{\mathrm{C}} \varphi$ holds, then $\mathrm{M}, s \models^{\mathrm{C}} \varphi$ for all $\mathrm{M} \in \mathbf{M}^{\mathrm{C}}$ with state s. Hence $\mathrm{M}, s \models^{\mathrm{C}} \neg\mathbf{Goal}_i\varphi$ for all $\mathrm{M} \in \mathbf{M}^{\mathrm{C}}$ and their states s, and thus $\models^{\mathrm{C}} \neg\mathbf{Goal}_i\varphi$.
6. This item is easily shown by selecting an appropriate contingency φ and an arbitrary tautology ψ, such that for certain M and s holds that $\mathrm{M}, s \models^{\mathrm{C}} \mathbf{Goal}_i\varphi$. For then $\mathrm{M}, s \models^{\mathrm{C}} (\varphi \rightarrow \psi) \wedge \mathbf{Goal}_i\varphi$ while — by the previous item — $\mathrm{M}, s \not\models^{\mathrm{C}} \mathbf{Goal}_i\psi$.
7. Item 7 is proved similarly to item 6.

⊠

6.26. Proposition. *For all $i \in \mathrm{A}$, $\alpha, \beta \in \mathrm{Ac}$ and $\varphi \in \mathrm{L}$ we have:*

1. $\models^{\mathrm{C}} \mathbf{Intend}_i(\alpha, \varphi) \rightarrow \langle\mathrm{do}_i(\mathtt{commit\_to}\,\alpha)\rangle\top$
2. $\models^{\mathrm{C}} \langle\mathrm{do}_i(\mathtt{commit\_to}\,\alpha)\rangle\top \leftrightarrow \langle\mathrm{do}_i(\mathtt{commit\_to}\,\alpha)\rangle\mathbf{Committed}_i\alpha$
3. $\models^{\mathrm{C}} \mathbf{Committed}_i\alpha \rightarrow \neg\mathbf{A}_i\mathtt{commit\_to}\,\beta$
4. $\models^{\mathrm{C}} [\mathrm{do}_i(\mathtt{commit\_to}\,\alpha)]\neg\mathbf{A}_i\mathtt{commit\_to}\,\beta$
5. $\models^{\mathrm{C}} \mathbf{Committed}_i\alpha \leftrightarrow \langle\mathrm{do}_i(\mathtt{uncommit}\,\alpha)\rangle\neg\mathbf{Committed}_i\alpha$
6. $\models^{\mathrm{C}} \mathbf{Intend}_i(\alpha, \varphi) \rightarrow \neg\mathbf{A}_i\mathtt{uncommit}\,\alpha$
7. $\models^{\mathrm{C}} \mathbf{A}_i\mathtt{uncommit}\,\alpha \leftrightarrow \mathbf{K}_i\mathbf{A}_i\mathtt{uncommit}\,\alpha$
8. $\models^{\mathrm{C}} \mathbf{Committed}_i\alpha \wedge \neg\mathbf{Can}_i(\alpha, \top) \rightarrow \mathbf{Can}_i(\mathtt{uncommit}\,\alpha, \neg\mathbf{Committed}_i\alpha)$

Proof: We show the second, third, fourth, seventh and eight item; the other ones follow directly from the respective definitions. Let $\mathrm{M} \in \mathbf{M}^{\mathrm{C}}$ with state s, and $i \in \mathrm{A}$, $\alpha, \beta \in \mathrm{Ac}$ be arbitrary.

2. Let $\mathrm{M}, s \models^{\mathrm{C}} \langle\mathrm{do}_i(\mathtt{commit\_to}\,\alpha)\rangle\top$ and let $\mathrm{M}', s = \mathtt{r}^{\mathrm{C}}(i, \mathtt{commit\_to}\,\alpha)(\mathrm{M}, s)$. We have to show that $\mathrm{M}', s \models^{\mathrm{C}} \mathbf{Committed}_i\alpha$, i.e. we have to show that $\forall s' \in [s]_{\mathrm{R}'(i)} \exists\alpha_1 \in \mathrm{CR}^{\mathrm{C}}_{\mathrm{M}'}(i, \alpha, s') \exists\alpha_2 \in \mathrm{Agenda}'(i, s')(\mathrm{Prefix}(\mathrm{normalised}(\alpha_1, \alpha_2)))$. A close inspection of Definition 6.19 and Figure 6.1 shows that for all $s' \in [s]_{\mathrm{R}(i)} = [s]_{\mathrm{R}'(i)}$ holds that $\mathrm{Agenda}'(i, s')$ contains $\mathrm{normalised}(\mathrm{CR}^{\mathrm{C}}_{\mathrm{M}}(i, \alpha, s'))$. Now for $\alpha \in \mathrm{Ac}$, $\mathrm{CR}^{\mathrm{C}}_{\mathrm{M}}(i, \alpha, s') = \mathrm{CR}^{\mathrm{C}}_{\mathrm{M}'}(i, \alpha, s')$, since α is treated identically in both M and M'. Hence for all $s' \in [s]_{\mathrm{R}'(i)}$ it holds that $\mathrm{normalised}(\mathrm{CR}^{\mathrm{C}}_{\mathrm{M}'}(i, \alpha, s')) \in \mathrm{Agenda}'(i, s')$, which implies that $\mathrm{M}', s \models^{\mathrm{C}} \mathbf{Committed}_i\alpha$. Thus $\mathrm{M}, s \models^{\mathrm{C}} \langle\mathrm{do}_i(\mathtt{commit\_to}\,\alpha)\mathbf{Committed}_i\alpha$, which suffices to conclude that item 2 holds.

3. If $\mathrm{M}, s \models^{\mathrm{C}} \mathbf{Committed}_i\alpha$ then, by Definition 6.22, we have that $\mathrm{Agenda}(i, s) \neq \emptyset$. Hence, by Definition 6.19, $\mathrm{M}, s \models^{\mathrm{C}} \neg\mathbf{A}_i\mathtt{commit\_to}\,\beta$.
4. If $\mathrm{r}^{\mathrm{C}}(i, \mathtt{commit\_to}\,\alpha)(\mathrm{M}, s) = \emptyset$ then $\mathrm{M}, s \models^{\mathrm{C}} [\mathrm{do}_i(\mathtt{commit\_to}\,\alpha)]\neg\mathbf{A}_i\mathtt{commit\_to}\,\beta$ is trivially true. Else $\mathrm{M}, s \models^{\mathrm{C}} \langle\mathrm{do}_i(\mathtt{commit\_to}\,\alpha)\rangle\mathbf{Committed}_i\alpha$ by item 2 of this proposition, and, by item 3, this implies $\mathrm{M}, s \models^{\mathrm{C}} \langle\mathrm{do}_i(\mathtt{commit\_to}\,\alpha)\rangle\neg\mathbf{A}_i\mathtt{commit\_to}\,\beta$, which suffices to conclude item 4.
7. Suppose $\mathrm{M}, s \models^{\mathrm{C}} \mathbf{A}_i\mathtt{uncommit}\,\alpha$. This implies that $\mathrm{M}, s \models^{\mathrm{C}} \neg\mathbf{Intend}_i(\alpha, \varphi)$, for all $\varphi \in \mathrm{C}(i, s)$. That is, $\mathrm{M}, s \models^{\mathrm{C}} \neg\mathbf{Can}_i(\alpha, \varphi) \vee \neg\mathbf{K}_i\mathbf{Goal}_i\varphi$ for all $\varphi \in \mathrm{C}(i, s)$. But by the introspective properties of knowledge the latter implies that $\mathrm{M}, s \models^{\mathrm{C}} \mathbf{K}_i\neg\mathbf{Can}_i(\alpha, \varphi) \vee \mathbf{K}_i\neg\mathbf{K}_i\mathbf{Goal}_i\varphi$, for all $\varphi \in \mathrm{C}(i, s)$. Hence $\mathrm{M}, s \models^{\mathrm{C}} \mathbf{K}_i(\neg\mathbf{Can}_i(\alpha, \varphi) \vee \neg\mathbf{K}_i\mathbf{Goal}_i\varphi)$, for all $\varphi \in \mathrm{C}(i, s)$, and thus for all $s' \in [s]_{\mathrm{R}(i)}$ it holds that $\mathrm{M}, s' \models^{\mathrm{C}} \neg\mathbf{Intend}_i(\alpha, \varphi)$ for all $\varphi \in \mathrm{C}(i, s)$, and thereby also $\mathrm{M}, s' \models^{\mathrm{C}} \mathbf{A}_i\mathtt{uncommit}\,\alpha$. Thus $\mathrm{M}, s \models^{\mathrm{C}} \mathbf{K}_i\mathtt{uncommit}\,\alpha$, which suffices to conclude that item 7 indeed holds.
8. Suppose $\mathrm{M}, s \models^{\mathrm{C}} \mathbf{Committed}_i\alpha \wedge \neg\mathbf{Can}_i(\alpha, \top)$. Then $\mathrm{M}, s \models^{\mathrm{C}} \neg\mathbf{Can}_i(\alpha, \varphi)$ for all $\varphi \in \mathrm{L}$, and thus $\mathrm{M}, s \models^{\mathrm{C}} \neg\mathbf{Intend}_i(\alpha, \varphi)$ for all $\varphi \in \mathrm{C}(i, s)$. Then, by definition of c^{C}, $\mathrm{M}, s \models^{\mathrm{C}} \mathbf{A}_i\mathtt{uncommit}\,\alpha$, and, by the previous item, $\mathrm{M}, s \models^{\mathrm{C}} \mathbf{K}_i\mathbf{A}_i\mathtt{uncommit}\,\alpha$. Also, $\mathrm{M}, s \models^{\mathrm{C}} \mathbf{Committed}_i\alpha$ implies $\mathrm{M}, s \models^{\mathrm{C}} \mathbf{K}_i\mathbf{Committed}_i\alpha$ by Proposition 6.27(1), and, by item 5 of this proposition, $\mathrm{M}, s \models^{\mathrm{C}} \mathbf{K}_i\langle\mathrm{do}_i(\mathtt{uncommit}\,\alpha)\rangle\neg\mathbf{Committed}_i\alpha$. Thus $\mathrm{M}, s \models^{\mathrm{C}} \mathbf{K}_i\langle\mathrm{do}_i(\mathtt{uncommit}\,\alpha)\rangle\neg\mathbf{Committed}_i\alpha \wedge \mathbf{K}_i\mathbf{A}_i\mathtt{uncommit}\,\alpha$. This implies that $\mathrm{M}, s \models^{\mathrm{C}} \mathbf{Can}_i(\mathtt{uncommit}\,\alpha, \neg\mathbf{Committed}_i\alpha)$, which suffices to conclude item 8.

⊠

6.27. Proposition. *For all* $i \in \mathrm{A}$, $\alpha, \alpha_1, \alpha_2 \in \mathrm{Ac}$ *and all* $\varphi \in \mathrm{L}$ *we have:*

1. $\models^{\mathrm{C}} \mathbf{Committed}_i\alpha \to \mathbf{K}_i\mathbf{Committed}_i\alpha$
2. $\models^{\mathrm{C}} \mathbf{Committed}_i\mathtt{if}\,\varphi\,\mathtt{then}\,\alpha_1\,\mathtt{else}\,\alpha_2\,\mathtt{fi} \wedge \mathbf{K}_i\varphi \to \mathbf{Committed}_i(\mathtt{confirm}\,\varphi; \alpha_1)$
3. $\models^{\mathrm{C}} \mathbf{Committed}_i\mathtt{if}\,\varphi\,\mathtt{then}\,\alpha_1\,\mathtt{else}\,\alpha_2\,\mathtt{fi} \wedge \mathbf{K}_i\neg\varphi \to \mathbf{Committed}_i(\mathtt{confirm}\,\neg\varphi; \alpha_2)$
4. $\models^{\mathrm{C}} \mathbf{Committed}_i\mathtt{while}\,\varphi\,\mathtt{do}\,\alpha\,\mathtt{od} \wedge \mathbf{K}_i\varphi \to$
   $\mathbf{Committed}_i((\mathtt{confirm}\,\varphi; \alpha); \mathtt{while}\,\varphi\,\mathtt{do}\,\alpha\,\mathtt{od})$

Proof: We successively show all items. Let $\mathrm{M} \in \mathbf{M}^{\mathrm{C}}$ with state s and $\varphi \in \mathrm{L}$, $\alpha, \alpha_1, \alpha_2 \in \mathrm{Ac}$ be arbitrary.

1. Assume that $\mathrm{M}, s \models^{\mathrm{C}} \mathbf{Committed}_i\alpha$. By Definition 6.22 it then follows that $\forall s' \in [s]_{\mathrm{R}(i)}\exists\alpha_1 \in \mathrm{CR}^{\mathrm{C}}_{\mathrm{M}}(i, \alpha, s')\exists\alpha_2 \in \mathrm{Agenda}(i, s')(\mathrm{Prefix}(\mathrm{normalised}(\alpha_1), \alpha_2))$. Since $[s]_{\mathrm{R}(i)}$ is an equivalence class we have that $\forall s'' \in [s]_{\mathrm{R}(i)}\forall s' \in [s'']_{\mathrm{R}(i)}\exists\alpha_1 \in \mathrm{CR}^{\mathrm{C}}_{\mathrm{M}}(i, \alpha, s')\exists\alpha_2 \in \mathrm{Agenda}(i, s')(\mathrm{Prefix}(\mathrm{normalised}(\alpha_1), \alpha_2))$, which implies $\mathrm{M}, s'' \models^{\mathrm{C}} \mathbf{Committed}_i\alpha$ for all $s'' \in [s]_{\mathrm{R}(i)}$, and thus $\mathrm{M}, s \models^{\mathrm{C}} \mathbf{K}_i\mathbf{Committed}_i\alpha$.
2. Assume that $\mathrm{M}, s \models^{\mathrm{C}} \mathbf{Committed}_i\mathtt{if}\,\varphi\,\mathtt{then}\,\alpha_1\,\mathtt{else}\,\alpha_2\,\mathtt{fi} \wedge \mathbf{K}_i\varphi$. By definition of $\mathrm{CR}^{\mathrm{C}}_{\mathrm{M}}$ and CS we have $\mathrm{CR}^{\mathrm{C}}_{\mathrm{M}}(i, \mathtt{if}\,\varphi\,\mathtt{then}\,\alpha_1\,\mathtt{else}\,\alpha_2\,\mathtt{fi}, s') = \mathrm{CR}^{\mathrm{C}}_{\mathrm{M}}(i, \mathtt{confirm}\,\varphi; \alpha_1, s')$ for all $s' \in [s]_{\mathrm{R}(i)}$. Hence it follows that $\exists\beta_1 \in \mathrm{CR}^{\mathrm{C}}_{\mathrm{M}}(i, \mathtt{if}\,\varphi\,\mathtt{then}\,\alpha_1\,\mathtt{else}\,\alpha_2\,\mathtt{fi}, s')\exists\beta_2 \in$

$\mathrm{Agenda}(i,s')(\mathrm{Prefix}(\mathrm{normalised}(\beta_1),\beta_2))$ implies $\exists\beta_1 \in \mathrm{CR}^{\mathrm{C}}_{\mathrm{M}}(i,(\mathtt{confirm}\,\varphi;\alpha_1),s')\exists\beta_2 \in \mathrm{Agenda}(i,s')(\mathrm{Prefix}(\mathrm{normalised}(\beta_1),\beta_2))$ for all $s' \in [s]_{\mathrm{R}(i)}$. Then it is indeed the case that from $\mathrm{M},s \models^{\mathrm{C}} \mathbf{Committed}_i\mathtt{if}\,\varphi\,\mathtt{then}\,\alpha_1\,\mathtt{else}\,\alpha_2\,\mathtt{fi}$ it follows that $\mathrm{M},s \models^{\mathrm{C}} \mathbf{Committed}_i(\mathtt{confirm}\,\varphi;\alpha_1)$, which suffices to conclude this item.

3. This item is completely analogous to the previous one.
4. From the definition of $\mathrm{CR}^{\mathrm{C}}_{\mathrm{M}}$ and CS it follows that in the case that $\mathrm{M},s \models^{\mathrm{C}} \varphi$, $\mathrm{CR}^{\mathrm{C}}_{\mathrm{M}}(i,\mathtt{while}\,\varphi\,\mathtt{do}\,\alpha\,\mathtt{od},s) = \mathrm{CR}^{\mathrm{C}}_{\mathrm{M}}(i,(\mathtt{confirm}\,\varphi;\alpha);\mathtt{while}\,\varphi\,\mathtt{do}\,\alpha\,\mathtt{od},s)$. By a similar argument as the one given in the proof of item 2 one concludes that $\mathrm{M},s \models^{\mathrm{C}} \mathbf{Committed}_i\mathtt{while}\,\varphi\,\mathtt{do}\,\alpha\,\mathtt{od}\wedge\mathbf{K}_i\varphi \rightarrow \mathbf{Committed}_i((\mathtt{confirm}\,\varphi;\alpha);\mathtt{while}\,\varphi\,\mathtt{do}\,\alpha\,\mathtt{od})$, which concludes item 4.

⊠

6.28. Proposition. *For all* $i \in \mathrm{A}$ *and* $\alpha_1,\alpha_2 \in \mathrm{Ac}$ *we have:*

- $\models^{\mathrm{C}} \mathbf{Committed}_i(\alpha_1;\alpha_2) \rightarrow \mathbf{Committed}_i\alpha_1$
- $\models^{\mathrm{C}}\langle \mathrm{do}_i(\mathtt{commit\_to}\,(\alpha_1;\alpha_2))\rangle\top \rightarrow$
  $\langle \mathrm{do}_i(\mathtt{commit\_to}\,(\alpha_1;\alpha_2))\rangle\mathbf{K}_i[\mathrm{do}_i(\alpha_1)]\mathbf{Committed}_i\alpha_2$

Proof: Let $\mathrm{M} \in \mathbf{M}^{\mathrm{C}}$ with state s, and $i \in \mathrm{A}$ and $\alpha_1,\alpha_2 \in \mathrm{Ac}$ be arbitrary.

- Let $\mathrm{M},s \models^{\mathrm{C}} \mathbf{Committed}_i\alpha_1;\alpha_2$, i.e. for all $s' \in [s]_{\mathrm{R}(i)}$ some $\beta_1 \in \mathrm{CR}^{\mathrm{C}}_{\mathrm{M}}(i,\alpha_1;\alpha_2,s')$ and $\beta_2 \in \mathrm{Agenda}(i,s')$ exist such that $\mathrm{Prefix}(\mathrm{normalised}(\beta_1),\beta_2)$ holds. By definition of CS for sequentially composed actions it follows that $\mathrm{normalised}(\mathrm{CR}^{\mathrm{C}}_{\mathrm{M}}(i,\alpha_1;\alpha_2,s'))$ is a sequence $\gamma_1;(\gamma_2;(\ldots;(\gamma_{l-1};\gamma_l)\ldots))$ with $\gamma_j \in \mathrm{Ac_s}$ such that for some $k \in \mathbb{N}$ it holds that $\mathrm{normalised}(\mathrm{CR}^{\mathrm{C}}_{\mathrm{M}}(i,\alpha_1,s')) = \gamma_1;(\ldots;(\gamma_{k-1};\gamma_k)\ldots)$. By definition of Prefix it is obvious that whenever β_2 is such that $\mathrm{Prefix}(\gamma_1;(\gamma_2;(\ldots;(\gamma_{l-1};\gamma_l)\ldots)),\beta_2)$ holds also $\mathrm{Prefix}(\gamma_1;(\gamma_2;(\ldots;(\gamma_{k-1};\gamma_k)\ldots)),\beta_2)$ holds. But this implies that whenever for all $s' \in [s]_{\mathrm{R}(i)}$ some $\beta_1 \in \mathrm{CR}^{\mathrm{C}}_{\mathrm{M}}(i,\alpha_1;\alpha_2,s')$ and $\beta_2 \in \mathrm{Agenda}(i,s')$ exist such that $\mathrm{Prefix}(\mathrm{normalised}(\beta_1),\beta_2)$ holds, also some $\beta_1 \in \mathrm{CR}^{\mathrm{C}}_{\mathrm{M}}(i,\alpha_1,s')$ and $\beta_2 \in \mathrm{Agenda}(i,s')$ exist such that $\mathrm{Prefix}(\mathrm{normalised}(\beta_1),\beta_2)$ holds. The latter suffices to conclude that $\mathrm{M},s \models^{\mathrm{C}} \mathbf{Committed}_i\alpha_1$. Thus if $\mathrm{M},s \models^{\mathrm{C}} \mathbf{Committed}_i\alpha_1;\alpha_2$ then $\mathrm{M},s \models^{\mathrm{C}} \mathbf{Committed}_i\alpha_1$, which concludes the proof of this item.
- Let $\mathrm{M},s \models^{\mathrm{C}} \langle \mathrm{do}_i(\mathtt{commit\_to}\,(\alpha_1;\alpha_2))\rangle\top$ and $\mathrm{M}',s = \mathbf{r}^{\mathrm{C}}(i,\mathtt{commit\_to}\,(\alpha_1;\alpha_2))(\mathrm{M},s)$. Let $s' \in [s]_{\mathrm{R}'(i)} = [s]_{\mathrm{R}(i)}$. An inspection of Figure 6.1 learns that for all states s'' in the epistemic equivalence class of $\mathbf{r}^{\mathrm{C}}(i,\mathrm{CR}^{\mathrm{C}}_{\mathrm{M}}(i,\alpha_1,s'))(\mathrm{M},s')$ holds that $\mathrm{CR}^{\mathrm{C}}_{\mathrm{M}}(i,\alpha_2,s'')$ is an element of i's agenda (note that $\mathrm{normalised}(\mathrm{CR}^{\mathrm{C}}_{\mathrm{M}}(i,\alpha_1,s'))$ is indeed a prefix of $\mathrm{normalised}(\mathrm{CR}^{\mathrm{C}}_{\mathrm{M}}(i,\alpha_1;\alpha_2,s'))$). Then $\mathrm{M},s' \models^{\mathrm{C}} [\mathrm{do}_i(\alpha_1)]\mathbf{Committed}_i\alpha_2$ and thus $\mathrm{M},s \models^{\mathrm{C}} \mathbf{K}_i[\mathrm{do}_i(\alpha_1)]\mathbf{Committed}_i\alpha_2$.

⊠

Chapter 7

Conclusions and future work

And when you feel you're near the end
And what once burned so bright is growing dim
And when you see what's been achieved
Is there a feeling that you've been deceived?

David Gilmour, *'Near the End'*.

In this brief chapter we globally reflect on the picture of agents as it is emerging from this thesis, and in particular compare this picture with the informal description presented in Chapter 1. As an example of how the formal machinery developed in the previous chapters could be put to work we present a sketchy specification of an (artificial) information agent. To conclude we summarise the main contributions of this thesis, and indicate open problems and opportunities for future research.

7.1 What's an agent, anyway?

The agents formalised in this thesis act in the world, and interact with the world and with other agents, both on a physical and on a mental level. They reason about their own, and other agents' acts, motives and information. The agents are rational, both with respect to their information, i.e. knowledge and various kinds of belief, and with respect to their acts, in particular when these acts are either nondeterministic or non-mundane. They are autonomous, in that they may themselves decide to adopt goals and to make or undo certain commitments, and they are social when communicating with other agents. Lastly, they have the possibility to acquire information through observations, communication or reasoning by default, and may use this information, for instance to reflect on their goals and commitments.

7.2 Modelling rational agents

To give the reader an impression as to how the machinery developed in the previous chapters could be used in the formal specification of artificial agents, we present an example specification of a fairly realistic information agent. This agent is similar to, and inspired by, the software agents implemented at the MIT Media Lab [88, 89, 90]. The language in which this specification is formulated is a combination of the languages presented in Chapter 5 and Chapter 6, where it has to be remarked that we take a somewhat liberal view on this. That is, we for example allow other formulae other than purely propositional ones to occur as the argument of an informative action, and we allow wishes to range over more general formulae than the ones considered in Chapter 6. The modifications necessary to semantically account for these changes in the syntax are — as far as we can see — fairly obvious and straightforwardly implementable.

7.1. Example. Consider an intelligent information agent that is assisting some user with his/her information management. The agent either observes the informational needs of the user or is told by the user what these needs consist of. The wishes of the user are the agent's commands, i.e. if the agent either observes or is told one of the user's wishes then it knows that this wish becomes one of its own. Starting from the wishes of the user, the agent infers its goals, and consecutively tries to come to a commitment to an appropriate action. This commitment is then communicated back to the user who decides whether the agent indeed has to live up to this commitment.

Now suppose the user wants to know whether some file is present on his/her disk, but s/he does not have direct access to this data. The agent however knows that it has the ability to find out whether this file is present, while it furthermore knows that the user depends on the agent for information on the presence of this file. In addition, the agent knows that it is ready to make commitments since it is as of yet not committed to any action whatsoever. Using the symbol u to represent the user, i to represent the agent, and f to represent the proposition 'the file is present on the disk', this situation is formally specified by the following four formulae. These formulae, as well as the other ones appearing in this example, are tagged to indicate their status: formulae tagged with an 'A' are assumptions underlying this specific example, formulae tagged with an 'L' are laws that are universally valid, and formulae tagged with a 'C' are logical consequences of the assumptions and the laws.

- $\mathbf{B}_i^c\mathbf{W}_u\varphi \rightarrow \mathbf{K}_i\mathbf{W}_i\varphi$ for all φ (A)
- $\mathbf{K}_i\mathbf{A}_i\texttt{observe}\,f$ (A)
- $\mathbf{K}_i(\mathbf{D}_{u,i}f \wedge \mathbf{D}_{u,i}\neg f)$ (A)
- $\mathbf{K}_i\neg\mathbf{Committed}_i\alpha$ for all α (A)

The user on its turn knows that s/he wants to know whether the file is present on his/her disk while not having any information of this kind at this moment. Furthermore, the user knows that the agent is willing to accept the user as an authority on any of his/her wishes. The attributes of the user are thus formalised:

- $\neg \mathbf{A}_u \texttt{observe}\, f$ (A)
- $\mathbf{K}_u(\mathbf{W}_u\mathbf{Bwhether}^k_u f \wedge \mathbf{Ignorant}^d_u f)$ (A)
- $\mathbf{K}_u(\mathbf{D}_{i,u}\mathbf{W}_u\varphi)$ for all φ (A)

Since the user knows of his/her wish to know whether the file is present on the disk and furthermore the agent accepts the user as an authority on his/her wishes, the user now has the opportunity to inform the agent of this wish. Assuming that the agent did not have any information on the wishes of the user prior to communication it will indeed accept the user's wish. The following formula, which follows from the validity given in item 6 of Proposition 5.30 and hence itself would be valid in the new semantics for the combined language, formalises this fact.

- $\mathbf{K}_u(\mathbf{W}_u\mathbf{Bwhether}^k_u f \wedge \mathbf{D}_{i,u}\mathbf{W}_u\mathbf{Bwhether}^k_u f) \wedge \mathbf{Ignorant}^d_i\mathbf{W}_u\mathbf{Bwhether}^k_u f \rightarrow$
  $\langle \mathrm{do}_u(\texttt{inform}\,(\mathbf{W}_u\mathbf{Bwhether}^k_u f, i))\rangle \mathbf{B}^c_i\mathbf{W}_u\mathbf{Bwhether}^k_u f$ (L)

As a result of being told that the user wants to know whether the file is present, the agent communicationally believes that it wishes the user to be in the possession of this information. Using the first equivalence given in the specification of the agent this implies that the agent knows that it wishes the user to know whether the file is present after it has been informed of the user's wish.

- $\mathbf{K}_u(\mathbf{W}_u\mathbf{Bwhether}^k_u f \wedge \mathbf{D}_{i,u}\mathbf{W}_u\mathbf{Bwhether}^k_u f) \wedge \mathbf{Ignorant}^d_i\mathbf{W}_u\mathbf{Bwhether}^k_u f \rightarrow$
  $\langle \mathrm{do}_u(\texttt{inform}\,(\mathbf{W}_u\mathbf{Bwhether}^k_u f, i))\rangle \mathbf{K}_i\mathbf{W}_i\mathbf{Bwhether}^k_u f$ (C)

By assumption, the agent knows that it has the ability to observe whether the file is present on the user's disk. Since observations are realisable (cf. item 2 of Proposition 5.27) the agent furthermore knows that it has the opportunity to observe the presence of the file. According to Definition 3.17 this implies that the agent knows that it has the practical possibility to do so.

- $\mathbf{K}_i\langle \mathrm{do}_i(\texttt{observe}\, f)\rangle \mathbf{Bwhether}^o_i f$ (L)
- $\mathbf{K}_i\mathbf{A}_i\texttt{observe}\, f$ (A)
- $\mathbf{Can}_i(\texttt{observe}\, f, \mathbf{Bwhether}^o_i f)$ (C)

The agent knows that if it observed the file to be present, it may successfully transfer this information to its user, which after all depends on the agent for information of this kind. Analogously, if the agent observed that the file is not on the disk, it has the reliable opportunity to tell the user that this is the case. Furthermore, the agent knows that having observational belief on the presence of the file implies having the ability to communicate this fact. These properties are formalised in the following four formulae. The first two of these formulae follow directly from the validity presented in

item 6 of Proposition 5.30 and are thus themselves valid. The last two formulae are straightforwardly seen to be valid when using Definition 5.37.

- $\mathbf{K}_i(\mathbf{B}_i^o f \wedge \mathbf{D}_{u,i} f \wedge \mathbf{Ignorant}_u^d f \rightarrow \langle \mathrm{do}_i(\mathtt{inform}\,(f,u))\rangle \mathbf{Heard}_u f)$ (L)
- $\mathbf{K}_i(\mathbf{B}_i^o \neg f \wedge \mathbf{D}_{u,i} \neg f \wedge \mathbf{Ignorant}_u^d f \rightarrow \langle \mathrm{do}_i(\mathtt{inform}\,(\neg f,u))\rangle \mathbf{Heard}_u \neg f)$ (L)
- $\mathbf{K}_i(\mathbf{B}_i^o f \rightarrow \mathbf{A}_i \mathtt{inform}\,(f,u))$ (L)
- $\mathbf{K}_i(\mathbf{B}_i^o \neg f \rightarrow \mathbf{A}_i \mathtt{inform}\,(\neg f,u))$ (L)

Assuming that the agent knows that the user does not have any information whatsoever on the presence of the file, and using that the agent knows that the user depends on it for information on the presence of the file on his/her disk, the first two of the formulae given directly above can be simplified to:

- $\mathbf{K}_i(\mathbf{B}_i^o f \rightarrow \langle \mathrm{do}_i(\mathtt{inform}\,(f,u))\rangle \mathbf{Heard}_u f)$ (C)
- $\mathbf{K}_i(\mathbf{B}_i^o \neg f \rightarrow \langle \mathrm{do}_i(\mathtt{inform}\,(\neg f,u))\rangle \mathbf{Heard}_u \neg f)$ (C)

Using the equivalence given in item 3 of Proposition 3.5 we can conclude from these formulae that the agent knows that if it observationally believes whether the file is present on the disk then performing a conditional composition that either amounts to telling the user that the file is present on the disk or that it is not, dependent on the information that the agent itself acquired through observation, will result in the user communicationally believing whether the file is indeed present. By an analogous line of reasoning and using the equivalence given in item 4 of Proposition 3.6, it follows that the agent knows that if it observationally believes that the file is present then it has the ability to perform this conditionally composed action.

- $\mathbf{K}_i(\mathbf{Bwhether}_i^o f \rightarrow$
  $\langle \mathrm{do}_i(\mathtt{if}\,\mathbf{B}_i^o f\,\mathtt{then}\,\mathtt{inform}\,(f,u)\,\mathtt{else}\,\mathtt{inform}\,(\neg f,u)\,\mathtt{fi})\rangle \mathbf{Bwhether}_u^c f)$ (C)
- $\mathbf{K}_i(\mathbf{Bwhether}_i^o f \rightarrow \mathbf{A}_i(\mathtt{if}\,\mathbf{B}_i^o f\,\mathtt{then}\,\mathtt{inform}\,(f,u)\,\mathtt{else}\,\mathtt{inform}\,(\neg f,u)\,\mathtt{fi}))$ (C)

Since the agent knows that an observation on the presence of the file on the user's disk results in it observationally believing whether this is the case, it also knows that sequentially composing the observation on the presence of the file with the conditionally encapsulated communication with the user constitutes a correct and feasible plan to transfer the desired information to its user.

- $\mathbf{K}_i(\langle \mathrm{do}_i(\mathtt{observe}\ f;$
  $\mathtt{if}\,\mathbf{B}_i^o f\,\mathtt{then}\,\mathtt{inform}\,(f,u)\,\mathtt{else}\,\mathtt{inform}\,(\neg f,u)\,\mathtt{fi})\rangle \mathbf{Bwhether}_u^c f)$ (C)
- $\mathbf{K}_i \mathbf{A}_i(\mathtt{observe}\ f; \mathtt{if}\,\mathbf{B}_i^o f\,\mathtt{then}\,\mathtt{inform}\,(f,u)\,\mathtt{else}\,\mathtt{inform}\,(\neg f,u)\,\mathtt{fi})$ (C)
- $\mathbf{Can}_i(\mathtt{observe}\ f; \mathtt{if}\,\mathbf{B}_i^o f\,\mathtt{then}\,\mathtt{inform}\,(f,u)\,\mathtt{else}\,\mathtt{inform}\,(\neg f,u)\,\mathtt{fi}, \mathbf{Bwhether}_u^c f)$ (C)

Let us denote the correct and feasible action that the agent came up with by `plan`, i.e. $\mathtt{plan} \triangleq \mathtt{observe}\ f; \mathtt{if}\,\mathbf{B}_i^o f\,\mathtt{then}\,\mathtt{inform}\,(f,u)\,\mathtt{else}\,\mathtt{inform}\,(\neg f,u)\,\mathtt{fi}$.

Slightly deviating from the original definition of the implementability operator, we assume that the practical possibility to bring about some proposition suffices to conclude that the proposition is implementable. Basically this comes down to declaring a formula

to be implementable if the agent has the practical possibility to perform an *arbitrary* (sequence of) action(s) — rather than a sequence of *atomic* actions as demanded in Definition 6.9 — that results in the formula being true.

- $\mathbf{PracPoss}_i(\alpha, \varphi) \rightarrow \Diamond_i \varphi$ for all α and all φ (A/L)

Since the agent (knows that it) has the practical possibility to bring it about that the user communicationally believes whether the file is present on the disk by performing its plan, it follows that (the agent knows that) truth of the latter formula is indeed implementable:

- $\mathbf{K}_i \Diamond_i \mathbf{Bwhether}^c_u f$ (C)

The combination of the implementability of the user communicationally believing whether the file is present with the unfulfilledness of this proposition suffices according to the formula

- $\mathbf{K}_i(\Diamond_i \mathbf{Bwhether}^c_u f \wedge \neg \mathbf{Bwhether}^c_u f \rightarrow \mathbf{A}_i \mathtt{select}\, \mathbf{Bwhether}^c_u f)$ (L)

which is a validity in the semantics given in Chapter 6, to conclude that the agent knows that it is able to select the proposition that the user communicationally believes whether the file is present on his/her disk:

- $\mathbf{K}_i \mathbf{A}_i \mathtt{select}\, \mathbf{Bwhether}^c_u f$ (C)

Since the agent's wishes are closed under logical consequence (cf. Proposition 6.8), and communicational beliefs are implied by knowledge (cf. Proposition 5.5), it follows from the fact that the agent knows that it wishes the user to know whether the file is present on his/her disk that the agent knows that it wishes the user to communicationally believe this.

- $\mathbf{K}_i \mathbf{W}_i \mathbf{Bwhether}^k_u f \rightarrow \mathbf{K}_i \mathbf{W}_i \mathbf{Bwhether}^c_u f$ (L)

Note that in Chapter 6 we did not consider the wish operator to range over doxastic formulae but only over formulae from L. When combining the language L^{I} from Chapter 5 with L^{C} from Chapter 6 it is however quite straightforward to let wishes range over formulae from L^{I}, including the doxastic ones.

Using the following formula, which is a validity in the semantics of Chapter 6, it follows that the agent has the opportunity to select this wish.

- $\mathbf{K}_i(\mathbf{W}_i \mathbf{Bwhether}^c_u f \rightarrow \langle \mathrm{do}_i(\mathtt{select}\, \mathbf{Bwhether}^c_u f) \rangle \mathbf{C}_i \mathbf{Bwhether}^c_u f)$ (L)

Since the agent knows that it has both the opportunity and ability — and hence the practical possibility — to select this wish, according to item 4 of Proposition 6.16 it follows that the agent knows that it can set the goal of the user communicationally believing whether the file is present on his/her disk:

- $\mathbf{K}_i(\langle \mathrm{do}_i(\mathtt{select}\, \mathbf{Bwhether}^c_u f) \rangle \mathbf{Goal}_i \mathbf{Bwhether}^c_u f)$ (C)

Assuming that the agent knows of its goals as soon as it has set them, it follows that the agent knows that it has the goal that the user is informed on the presence of the file

on the disk. The knowledge of this goal in combination with the agent's knowledge on its practical possibility to achieve this goal, suffices according to the following formula, which is a combination of the first two items of Proposition 6.26, to conclude that the agent may commit itself to its plan.

- $\mathbf{K}_i\mathbf{Goal}_i\mathbf{Bwhether}^c_u f \wedge \mathbf{Can}_i(\mathtt{plan}, \mathbf{Bwhether}^c_u f) \rightarrow$
  $\langle \mathrm{do}_i(\mathtt{commit\_to}\,(\mathtt{plan}))\rangle \mathbf{Committed}_i\mathtt{plan}$ (L)

Since the agent knows of its commitments (cf. item 1 of Proposition 6.27), it may now inform the user of its commitment to the plan. Assuming that the user is both ignorant on this fact and willing to accept the information of the agent, this transfer of information results in the user communicationally believing that the agent is committed. This is formalised in the following formula, which is a validity in the combined semantics for the languages of Chapter 5 and Chapter 6.

- $\mathbf{K}_i\mathbf{Committed}_i\mathtt{plan} \wedge \mathbf{D}_{u,i}\mathbf{Committed}_i\mathtt{plan} \wedge \mathbf{Ignorant}^d_u\mathbf{Committed}_i\mathtt{plan} \rightarrow$
  $\langle \mathrm{do}_i(\mathtt{inform}\,(\mathbf{Committed}_i\mathtt{plan}, u))\rangle \mathbf{B}^c_u\mathbf{Committed}_i\mathtt{plan}$ (L)

Since the agent knows both of its commitment to its plan and of the user's dependence on it for information on this commitment, the user indeed becomes informed of the agent's commitment:

- $\mathbf{B}^c_u\mathbf{Committed}_i\mathtt{plan}$ (C)

It is now up to the user whether the agent has to live up to its commitment and perform its plan.

7.3 Achievements

In this thesis we investigated the use and usability of modal logic as a tool to formalise rational agents. In doing so, we combined various modal logics, viz. epistemic, doxastic and dynamic logic, in one formal framework. This combination in itself is already considerably original, but even more so is the incorporation of ability as a first-class citizen. The latter allows us to reason with, and about, opportunity and ability as independent notions, which enhances the expressive power of the framework. Various interesting notions relating opportunity, ability, result and knowledge are defined in terms of the framework, which furthermore allows for a formalisation of philosophically interesting notions, like for example practical possibility. Sound and complete axiomatisations of two kinds of validity in the class of models of the basic formal system are given, the most remarkable feature of which is the use of infinitary proof rules. The proof that these axiomatisations are indeed sound and complete is a rather elaborate and fairly complex one. The first possible extension of the framework that we considered deals with nondeterminism of actions. Due to the presence of ability as a primitive, i.e. non-reducible, notion in our framework the need arises for non-standard formalisations of

nondeterminism. We presented two novel approaches, which are based on the unravelling of actions, and one which is an adaptation of the one proposed by Peleg in his Concurrent Propositional Dynamic Logic. The pragmatic, algorithmic character of the two approaches based on the unravelling of actions clearly shows their origin to be in computer science. The second extension of the basic framework provided a formalisation of intelligent information agents, i.e. agents of which the main task is to manage information. In defining these agents we presented a novel extension of the usual semantics for doxastic logic. This extended semantics allows for a classification of beliefs according to their credibility. To model the information acquisition of agents we introduced special, so-called informative actions. The semantics of these actions is based on a new paradigm for Propositional Dynamic Logic. This new paradigm generalises the standard one, in that actions may cause transformations of models in addition to transitions between states. Using this more general paradigm we succeeded in modelling informative actions in such a way that compliance with the AGM postulates for belief change is guaranteed. Another novel aspect of our modelling of intelligent information agents concerns the formalisation of supernormal defaults. We proposed to model these defaults by the epistemic notion of common possibility. This allows us to formalise defaults in their entirety within the formal framework, without resorting to additional formal tools. The last formalisation presented in this thesis concerns the agents' motivational attitudes. The formalisation as we present it deals with a broad range of motivational attitudes, both at the level of assertions, where we consider wishes and goals, and at the level of practitions, where we formalise commitments. In contrast with common practice, we did not define goals to be primitive, but in terms of selected wishes. Apart from being more acceptable from a philosophical and psychological point of view, this allowed us to avoid all kinds of problems that are well-known to plague formalisations of motivational attitudes. With regard to the agent's selections and commitments we considered both a static and a dynamic aspect, the latter of which is formalised using the generalised paradigm for Propositional Dynamic Logic that we previously introduced.

On the whole, we investigated the boundaries of expressiveness of modal logic when used to model rational agents. The results that we achieved in trying to extend the standard account of the various modal logics, viz. the extension of doxastic logic that allows one to classify beliefs according to credibility and the generalised paradigm for dynamic logic, combined with the easy adaptation of our framework to one's personal preferences, clearly shows the flexibility of our approach. Given the extensions proposed in this thesis, we are of the opinion that modal logic can indeed serve as a flexible and expressive yet intelligible formal tool in the analysis and specification of agents and agency. That is not to say that the road to genuine practical use of the formalism presented in this thesis might not be long and full of unexpected obstacles.

7.4 Future research

We foresee several directions in which the research captured in this thesis could proceed. The most obvious one is to further investigate, and eventually implement, the suggestions for possible extensions as made throughout this thesis. In addition to this, we feel that it could be worthwhile and interesting to come up with a first-order version of the logical systems presented here. This would on the one hand certainly involve its own specific problems, on the other hand it would enlarge expressiveness to a considerable extent. Another area for further research covers the introduction of temporal notions in the various formal systems. The incorporation of (implicit or explicit) notions of time and duration would probably be necessary to make the formalisms suitable for practical applications.

In addition to these extensions at the object-level, there is also a lot of research to be carried out at a higher level. In the first instance this high level research concerns the subjects that we explicitly mentioned not to be covered by this thesis. That is, research on the (meta-)logical properties of the various formal systems is still to be conducted. Topics like decidability and complexity have as of yet not been dealt with, and sound and complete axiomatisations for systems other than the most basic ones are lacking.

On a practical level we foresee two main topics of research. Firstly, in the spirit of the specification given in Example 7.1 one should try to formally specify and verify existing (software) agents. Secondly, it could be interesting to look at fragments of the various systems that could be fed to a theorem prover for automatic verification of formal specifications. An investigation of this kind would probably go hand in hand with the study of meta-logical properties like decidability and complexity.

It is our hope that the formal systems presented in this thesis will prove to be a fertile soil for further investigation, both from a theoretical and from a practical perspective.

Bibliography

[1] C.E. Alchourrón, P. Gärdenfors, and D. Makinson. On the logic of theory change: partial meet contraction and revision functions. *Journal of Symbolic Logic*, 50:510–530, 1985.

[2] K.R. Apt and G.D. Plotkin. Countable nondeterminism and random assignment. *Journal of the ACM*, 33(4):724–767, 1986.

[3] L. Åqvist. Deontic logic. In D.M. Gabbay and F. Guenthner, editors, *Handbook of Philosophical Logic*, volume 2, chapter 11, pages 605–714. D. Reidel, Dordrecht, 1984.

[4] I. Asimov. *The Complete Robot*. Doubleday, 1982.

[5] I. Asimov. *Robot Dreams*. Byron Preiss, 1986.

[6] I. Asimov. *Robot Visions*. Byron Preiss, 1990.

[7] J. Barwise. *The Situation in Logic*, volume 17 of *CSLI Lecture Notes*. CSLI, Stanford, 1989.

[8] J. Bates. The nature of characters in interactive worlds and the Oz project. In C.E. Loeffler, editor, *Virtual Realities: Anthology of Industry and Culture*, 1993.

[9] M.E. Bratman. *Intentions, Plans, and Practical Reason*. Harvard University Press, Cambridge, MA, 1987.

[10] M.A. Brown. On the logic of ability. *Journal of Philosophical Logic*, 17:1–26, 1988.

[11] M. Broy. A theory of nondeterminism, parallelism, communication and concurrency. *Theoretical Computer Science*, 45:1–61, 1986.

[12] R.W. Butler and G.B. Finelli. The infeasibility of experimental quantification of life-critical software reliability. In *Proceedings of the ACM SIGSOFT'91 Conference on Software for Critical Systems*, pages 66–76, 1991.

[13] R.W. Butler and G.B. Finelli. The infeasibility of quantifying the reliability of life-critical real-time software. *IEEE Transactions on Software Engineering*, 19(1):3–12, 1993.

[14] H.-N. Castañeda. The paradoxes of deontic logic: the simplest solution to all of them in one fell swoop. In Risto Hilpinen, editor, *New Studies in Deontic Logic*, pages 37–85. Reidel, Dordrecht, 1981.

[15] C. Castelfranchi. Personal communication.

[16] C. Castelfranchi. Guarantees for autonomy in cognitive agent architecture. In M. Wooldridge and N.R. Jennings, editors, *Intelligent Agents – Agent Theories, Architectures, and Languages*, volume 890 of *Lecture Notes in Computer Science (subseries LNAI)*, pages 56–70. Springer-Verlag, 1995.

[17] C. Castelfranchi, D. D'Aloisi, and F. Giacomelli. A framework for dealing with belief-goal dynamics. In M. Gori and G. Soda, editors, *Topics in Artificial Intelligence*, volume 992 of *Lecture Notes in Computer Science (subseries LNAI)*, pages 237–242. Springer-Verlag, 1995.

[18] L. Cavedon, L. Padgham, A. Rao, and E. Sonenberg. Revisiting rationality for agents with intentions. In X. Yao, editor, *Bridging the Gap: Proceedings of the Eight Australian Joint Conference on Artificial Intelligence*, pages 131–138. World Scientific, 1995.

[19] B.F. Chellas. *Modal Logic. An Introduction.* Cambridge University Press, Cambridge, 1980.

[20] P.R. Cohen and H.J. Levesque. Intention is choice with commitment. *Artificial Intelligence*, 42:213–261, 1990.

[21] *Communications of the ACM*, vol. 37, nr. 7. Special Issue on Intelligent Agents.

[22] A. Darwiche and J. Pearl. On the logic of iterated belief revision. In R. Fagin, editor, *Proceedings of the Fifth Conference on Theoretical Aspects of Reasoning about Knowledge (TARK'94)*, pages 5–23, Pacific Grove, CA, 1994. Morgan Kaufmann.

[23] J. W. de Bakker. *Mathematical Theory of Program Correctness.* Prentice-Hall, 1980.

[24] N. Dershowitz and J.-P. Jouannaud. Rewrite systems. In J. van Leeuwen, editor, *Handbook of Theoretical Computer Science*, volume B, pages 243–320. Elsevier, 1990.

[25] F. Dignum and B. van Linder. Modelling rational agents in a dynamic environment: Putting Humpty Dumpty together again. In J.L. Fiadeiro and P.-Y. Schobbens, editors, *Proceedings of the 2nd Workshop of the ModelAge Project*, pages 81–91, 1996.

[26] F. Dignum and B. van Linder. Modelling social agents in a dynamic environment: Making agents talk. Submitted, 1996.

[27] F. Dignum, J.-J.Ch. Meyer, R.J. Wieringa, and R. Kuiper. A modal approach to intentions, commitments and obligations: Intention plus commitment yields obligation. In M.A. Brown and J. Carmo, editors, *Deontic Logic, Agency and Normative Systems*, Springer Workshops in Computing, pages 80–97. Springer-Verlag, 1996.

[28] B. Dunin-Keplicz and A. Radzikowska. Epistemic approach to actions with typical effects. In C. Froidevaux and J. Kohlas, editors, *Symbolic and Quantitative Approaches to Reasoning and Uncertainty*, volume 946 of *Lecture Notes in Computer Science (subseries LNAI)*, pages 180–188. Springer-Verlag, 1995.

[29] D. Elgesem. *Action Theory and Modal Logic*. PhD thesis, Institute for Philosophy, University of Oslo, Oslo, Norway, 1993.

[30] E.A. Emerson. Temporal and modal logic. In J. van Leeuwen, editor, *Handbook of Theoretical Computer Science*, volume B, pages 995–1072. Elsevier, 1990.

[31] R. Fagin and J.Y. Halpern. Belief, awareness and limited reasoning. *Artificial Intelligence*, 34:39–76, 1988.

[32] J.L. Fiadeiro and P.-Y. Schobbens, editors. *Proceedings of the 2nd Workshop of the ModelAge Project*, 1996.

[33] L.N. Foner. What's an agent, anyway? A sociological case study. Technical report, MIT Media Laboratory, 1993.

[34] D. Gabbay. An irreflexivity lemma with applications to axiomatizations of conditions on linear frames. In U. Monnich, editor, *Aspects of Philosophical Logic*. D. Reidel, Dordrecht, 1981.

[35] D.M. Gabbay and F. Guenthner, editors. *Handbook of Philosophical Logic*, volume 2. D. Reidel, Dordrecht, 1984.

[36] L.F.T. Gamut. *Logic, Language and Meaning. Volume I: Introduction to Logic*. The University of Chicago Press, Chicago and London, 1991.

[37] L.F.T. Gamut. *Logic, Language and Meaning. Volume II: Intensional Logic and Logical Grammar*. The University of Chicago Press, Chicago and London, 1991.

[38] P. Gärdenfors. *Knowledge in Flux: Modeling the Dynamics of Epistemic States*. The MIT Press, Cambridge, Massachusetts and London, England, 1988.

[39] P. Gärdenfors, editor. *Belief Revision*. Cambridge University Press, 1992.

[40] G. Gazdar, G. Pullum, R. Carpenter, E. Klein, T. Hukari, and R. Levine. Category structures. *Computational Linguistics*, 14:1–19, 1988.

[41] R. Goldblatt. *Axiomatising the Logic of Computer Programming*, volume 130 of *LNCS*. Springer-Verlag, 1982.

[42] R. Goldblatt. The semantics of Hoare's iteration rule. *Studia Logica*, 41:141–158, 1982.

[43] R. Goldblatt. *Logics of Time and Computation*, volume 7 of *CSLI Lecture Notes*. CSLI, Stanford, 1992. Second edition.

[44] J. Halpern and J. Reif. The propositional dynamic logic of deterministic, well-structured programs. *Theoretical Computer Science*, 27:127–165, 1983.

[45] J.Y. Halpern and Y. Moses. A guide to completeness and complexity for modal logics of knowledge and belief. *Artificial Intelligence*, 54:319–379, 1992.

[46] D. Harel. Dynamic logic. In D.M. Gabbay and F. Guenthner, editors, *Handbook of Philosophical Logic*, volume 2, chapter 10, pages 497–604. D. Reidel, Dordrecht, 1984.

[47] D. Hilbert. Die Grundlegung der elementaren Zahlenlehre. *Mathematische Annalen*, 104:485–494, 1931.

[48] J. Hintikka. *Knowledge and Belief*. Cornell University Press, Ithaca, NY, 1962.

[49] C.A.R. Hoare. An axiomatic basis for computer programming. *Communications of the ACM*, 12:576–580, 1969.

[50] C.A.R. Hoare. *Communicating Sequential Processes*. Prentice-Hall International, 1985.

[51] W. van der Hoek. Systems for knowledge and beliefs. *Journal of Logic and Computation*, 3(2):173–195, 1993.

[52] W. van der Hoek, B. van Linder, and J.-J. Ch. Meyer. A logic of capabilities. Technical Report IR-330, Vrije Universiteit Amsterdam, July 1993.

[53] W. van der Hoek, B. van Linder, and J.-J. Ch. Meyer. Unravelling nondeterminism: On having the ability to choose. Technical Report RUU-CS-93-30, Utrecht University, September 1993.

[54] W. van der Hoek, B. van Linder, and J.-J. Ch. Meyer. A logic of capabilities. In A. Nerode and Yu. V. Matiyasevich, editors, *Proceedings of the Third International Symposium on the Logical Foundations of Computer Science (LFCS'94)*, volume 813 of *Lecture Notes in Computer Science*, pages 366–378. Springer-Verlag, 1994.

[55] W. van der Hoek, B. van Linder, and J.-J. Ch. Meyer. Unravelling nondeterminism: On having the ability to choose (extended abstract). In P. Jorrand and V. Sgurev, editors, *Proceedings of the Sixth International Conference on Artificial Intelligence: Methodology, Systems, Applications (AIMSA'94)*, pages 163–172. World Scientific, 1994.

[56] W. van der Hoek, B. van Linder, and J.-J. Ch. Meyer. Group knowledge isn't always distributed (neither is it always implicit). In M. Koppel and E. Shamir, editors, *Proceedings of BISFAI'95*, pages 191–200, 1995.

[57] W. van der Hoek, B. van Linder, and J.-J.Ch. Meyer. Modelling rational agents using modal logic. In *Proceedings of the First International Workshop on Decentralized Intelligent Multi Agent Systems (DIMAS'95)*, pages 215–224, 1995.

[58] J. Horty and Y. Shoham, Program Chairs. Reasoning about mental states: Formal theories & applications. Technical Report SS-93-05, AAAI Press, 1993. Papers from the 1993 AAAI Spring Symposium, Stanford CA.

[59] Z. Huang. Logics for belief dependence. In E. Börger, H. Kleine Büning, M.M. Richter, and W. Schönfeld, editors, *Computer Science Logic, 4th Workshop CSL'90*, volume 533 of *Lecture Notes in Computer Science*, pages 274–288. Springer-Verlag, 1991.

[60] G.E. Hughes and M.J. Cresswell. *An Introduction to Modal Logic*. Routledge, London, 1968.

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

[62] T.W.C. Huibers and B. van Linder. Formalising intelligent information retrieval agents. In F. Johnson, editor, *Proceedings of the 18th BCS IRSG Annual Colloquium on Information Retrieval Research*, pages 125–143, 1996.

[63] T.W.C. Huibers, B. van Linder, and P.D. Bruza. Een theorie voor het bestuderen van information retrieval modellen. In *Informatiewetenschap 1994: Wetenschappelijke Bijdragen aan de Derde StinfoN Conferentie*, pages 85–102. Stichting StinfoN, 1994.

[64] A.J.I. Jones. Practical reasoning, California-style: Some remarks on Shoham's agent-oriented programming. Medlar deliverable, 1993.

[65] S. Kanger. Law and logic. *Theoria*, 38, 1972.

[66] I. Kant. *Kritik der reinen Vernunft.* Thienemann, 1905. Reprint of the first edition of 1781.

[67] H. Katsumo and A.O. Mendelzon. On the difference between updating a knowledge base and revising it. In P. Gärdenfors, editor, *Belief revision*, pages 183–203. Cambridge University Press, 1992.

[68] A. Kenny. *Will, Freedom and Power.* Basil Blackwell, Oxford, 1975.

[69] J.W. Klop. Term rewriting systems. In S. Abramsky, D.M. Gabbay, and T.S.E. Maibaum, editors, *Handbook of Logic in Computer Science*, volume 2, pages 1–116. Oxford University Press, New York, 1992.

[70] D. Kozen and J. Tiuryn. Logics of programs. In J. van Leeuwen, editor, *Handbook of Theoretical Computer Science*, volume B, pages 789–840. Elsevier, 1990.

[71] S. Kraus and D. Lehmann. Knowledge, belief and time. *Theoretical Computer Science*, 58:155–174, 1988.

[72] S. Kripke. Semantic analysis of modal logic. *Zeitschrift für Mathematische Logik und Grundslagen der Mathematik*, 9:67–96, 1963.

[73] F. Kröger. Infinite proof rules for loops. *Acta Informatica*, 14:371–389, 1980.

[74] C. Krogh. Obligations in multiagent systems. In A. Aamodt and J. Komorowski, editors, *SCAI'95 – Fifth Scandinavian Conference on Artificial Intelligence*, pages 19–30. IOS Press, 1995.

[75] H. Kyburg. *Probability and the Logic of Rational Belief.* Wesleyan University Press, Middleton, Connecticut, 1961.

[76] H. Leblanc and W.A. Wisdom. *Deductive Logic*. Prentice Hall, 3 edition, 1993.

[77] D. Lehmann. Belief revision, revised. In C.S. Mellish, editor, *Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence (IJCAI'95)*, pages 1534–1540. Morgan Kaufmann, 1995.

[78] Y. Lespérance, H. Levesque, F. Lin, D. Marcu, R. Reiter, and R. Scherl. Foundations of a logical approach to agent programming. In M. Wooldridge, J.P. Müller, and M. Tambe, editors, *Intelligent Agents Volume II – Agent Theories, Architectures, and Languages*, volume 1037 of *Lecture Notes in Computer Science (subseries LNAI)*, pages 331–347. Springer-Verlag, 1996.

[79] V. Lesser, editor. *Proceedings of the First International Conference on Multi-Agent Systems (ICMAS'95)*. MIT Press, 1995.

[80] H. Levesque. A logic of implicit and explicit belief. In *Proceedings of the Fourth National Conference on Artificial Intelligence (AAAI'84)*, pages 198–202. The AAAI Press/The MIT Press, 1984.

[81] B. van Linder. A dynamic logic of iterated belief change. In X. Yao, editor, *Bridging the Gap: Proceedings of the Eight Australian Joint Conference on Artificial Intelligence*, pages 419–426. World Scientific, 1995.

[82] B. van Linder, W. van der Hoek, and J.-J. Ch. Meyer. Communicating rational agents. In B. Nebel and L. Dreschler-Fischer, editors, *KI-94: Advances in Artificial Intelligence*, volume 861 of *Lecture Notes in Computer Science (subseries LNAI)*, pages 202–213. Springer-Verlag, 1994.

[83] B. van Linder, W. van der Hoek, and J.-J. Ch. Meyer. Tests as epistemic updates. In A.G. Cohn, editor, *Proceedings of the 11th European Conference on Artificial Intelligence (ECAI'94)*, pages 331–335. John Wiley & Sons, 1994.

[84] B. van Linder, W. van der Hoek, and J.-J. Ch. Meyer. Actions that make you change your mind. In A. Laux and H. Wansing, editors, *Knowledge and Belief in Philosophy and Artificial Intelligence*, pages 103–146. Akademie Verlag, 1995.

[85] B. van Linder, W. van der Hoek, and J.-J. Ch. Meyer. The dynamics of default reasoning (extended abstract). In C. Froidevaux and J. Kohlas, editors, *Symbolic and Quantitative Approaches to Reasoning and Uncertainty*, volume 946 of *Lecture Notes in Computer Science (subseries LNAI)*, pages 277–284. Springer-Verlag, 1995. Full version to appear in *Data & Knowledge Engineering*.

[86] B. van Linder, W. van der Hoek, and J.-J. Ch. Meyer. Formalising motivational attitudes of agents: On preferences, goals and commitments. In M. Wooldridge, J.P. Müller, and M. Tambe, editors, *Intelligent Agents Volume II – Agent Theories, Architectures, and Languages*, volume 1037 of *Lecture Notes in Computer Science (subseries LNAI)*, pages 17–32. Springer-Verlag, 1996.

[87] B. van Linder, W. van der Hoek, and J.-J.Ch. Meyer. Seeing is believing – and so are hearing and jumping. In M. Gori and G. Soda, editors, *Topics in Artificial Intelligence*, volume 861 of *Lecture Notes in Computer Science (subseries LNAI)*, pages 402–413. Springer-Verlag, 1995. Extended version to appear in the *Journal of Logic, Language and Information*.

[88] P. Maes. Agents that reduce work and information overload. *Communications of the ACM*, 37(7):30–40, July 1994.

[89] P. Maes. Intelligent software. *Scientific American*, 273(3):66–68, September 1995. Special Issue on Key Technologies for the 21st Century.

[90] P. Maes and R. Kozierok. Learning interface agents. In *Proceedings of the Eleventh National Conference on Artificial Intelligence (AAAI'93)*, pages 459–465. The AAAI Press/The MIT Press, 1993.

[91] V.W. Marek and M. Truszczyński. *Nonmonotonic Logic*. Springer-Verlag, 1993.

[92] J.-J. Ch. Meyer. A different approach to deontic logic: Deontic logic viewed as a variant of dynamic logic. *Notre Dame Journal of Formal Logic*, 29:109–136, 1988.

[93] J.-J. Ch. Meyer. Free choice permissions and Ross's paradox: Internal vs external nondeterminism. In P. Dekker and M. Stokhof, editors, *Proceedings of the 8th Amsterdam Colloquium*, pages 367–380. Universiteit van Amsterdam, 1992.

[94] J.-J. Ch. Meyer and W. van der Hoek. A modal logic for nonmonotonic reasoning. In W. van der Hoek, J.-J. Ch. Meyer, Y.H. Tan, and C. Witteveen, editors, *Non-Monotonic Reasoning and Partial Semantics*, pages 37–77. Ellis Horwood, Chichester, 1992.

[95] J.-J. Ch. Meyer and W. van der Hoek. A default logic based on epistemic states. *Fundamentae Informatica*, 23(1):33–65, 1995.

[96] J.-J. Ch. Meyer and W. van der Hoek. *Epistemic Logic for AI and Computer Science*. Cambridge University Press, 1995.

[97] J.-J. Ch. Meyer and R.J. Wieringa. Deontic logic: A concise overview. In J.-J. Ch. Meyer and R.J. Wieringa, editors, *Deontic Logic in Computer Science*, chapter 1, pages 3–16. John Wiley & Sons, 1993.

[98] M. Miceli, A. Cesta, and P. Rizzo. Distributed artificial intelligence from a socio-cognitive standpoint: Looking at reasons for interaction. Technical report, Institute of Psychology, CNR, Rome, Italy, 1995.

[99] R.C. Moore. Reasoning about knowledge and action. Technical Report 191, SRI International, 1980.

[100] R.C. Moore. A formal theory of knowledge and action. In J.R. Hobbs and R.C. Moore, editors, *Formal Theories of the Commonsense World*, pages 319–358. Ablex, Norwood, NJ, 1985.

[101] D. Peleg. Concurrent dynamic logic. *Journal of the ACM*, 34(2):450–479, 1987.

[102] B. Penther. A dynamic logic of action. *Journal of Logic, Language and Information*, 3(3):169–210, 1994.

[103] D. Poole. A logical framework for default reasoning. *Artificial Intelligence*, 36:27–47, 1988.

[104] I. Pörn. *The Logic of Power*. Basil Blackwell, Oxford, 1970.

[105] I. Pörn. *Action Theory and Social Science*. Reidel, Dordrecht, 1977.

[106] V. Rantala. Impossible worlds semantics and logical omniscience. *Acta Philosophica Fennica*, 35:106–115, 1982.

[107] A. S. Rao. Means-end plan recognition – towards a theory of reactive recognition. In J. Doyle, E. Sandewall, and P. Torasso, editors, *Proceedings of the Fourth International Conference on Principles of Knowledge Representation and Reasoning (KR'94)*, pages 497–509. Morgan Kaufmann, 1994.

[108] A.S. Rao and M.P. Georgeff. Asymmetry thesis and side-effect problems in linear time and branching time intention logics. In J. Mylopoulos and R. Reiter, editors, *Proceedings of the Twelfth International Joint Conference on Artificial Intelligence (IJCAI'91)*, pages 498–504. Morgan Kaufmann, 1991.

[109] A.S. Rao and M.P. Georgeff. Modeling rational agents within a BDI-architecture. In J. Allen, R. Fikes, and E. Sandewall, editors, *Proceedings of the Second International Conference on Principles of Knowledge Representation and Reasoning (KR'91)*, pages 473–484. Morgan Kaufmann, 1991.

[110] A.S. Rao and M.P. Georgeff. A model-theoretic approach to the verification of situated reasoning systems. In R. Bajcsy, editor, *Proceedings of the Thirteenth International Joint Conference on Artificial Intelligence (IJCAI'93)*, pages 318–324. Morgan Kaufmann, 1993.

[111] J. Raz, editor. *Practical Reasoning*. Oxford Readings in Philosophy. Oxford University Press, 1978.

[112] R. Reiter. A logic for default reasoning. *Artificial Intelligence*, 13:81–132, 1980.

[113] R. Reiter and G. Crisculo. On interacting defaults. In *Proceedings of the Seventh International Joint Conference on Artificial Intelligence (IJCAI'81)*, pages 270–276. Morgan Kaufmann, 1981.

[114] D. Riecken. Intelligent agents. *Communications of the ACM*, 37(7):18–21, July 1994.

[115] A. Ross. Imperatives and logic. *Theoria*, 7:53–71, 1941.

[116] R.B. Scherl and H.J. Levesque. The frame problem and knowledge-producing actions. In *Proceedings of the Eleventh National Conference on Artificial Intelligence (AAAI'93)*, pages 689–694. The AAAI Press/The MIT Press, 1994. Submitted to *Artificial Intelligence*.

[117] K. Schütte. *Beweistheorie*. Springer-Verlag, Berlin-Göttingen-Heidelberg, 1960.

[118] K. Segerberg. Bringing it about. *Journal of philosophical logic*, 18:327–347, 1989.

[119] T. Selker. Coach: A teaching agent that learns. *Communications of the ACM*, 37(7):92–99, July 1994.

[120] Y. Shoham. Implementing the intentional stance. In R. Cummins and J. Pollock, editors, *Philosophy and AI: Essays at the Interface*. MIT Press, Cambridge, MA, 1991.

[121] Y. Shoham. Agent-oriented programming. *Artificial Intelligence*, 60:51–92, 1993.

[122] E. Spaan. *Complexity of Modal Logics*. PhD thesis, Universiteit van Amsterdam, 1993.

[123] P.A. Spruit. Henkin-style completeness proofs for Propositional Dynamic Logic. Manuscript.

[124] A. Tarski. A lattice-theoretical fixpoint theorem and its applications. *Pacific Journal of Mathematics*, 5:285–309, 1955.

[125] E. Thijsse. *Partial Logic and Knowledge Representation.* PhD thesis, Katholieke Universiteit Brabant, 1992.

[126] G. Tidhar. Personal communication.

[127] G. Tidhar, M. Slevestral, and C. Heinze. Modelling teams and team tactics in whole air mission modelling. In G. F. Forsyth and M. Ali, editors, *Proceedings of the Eighth International Conference on Industrial and Engineering Applications of Artificial Intelligence and Expert Systems (IEA/AIE'95)*, pages 373–381. Gordon and Breach Publishers, 1995.

[128] Web site of the Formal Methods Team of NASA Langley. URL: http://atb-www.larc.nasa.gov/cgi-bin/fm.cgi.

[129] M. Wooldridge and N. R. Jennings. Intelligent agents: Theory and practice. *The Knowledge Engineering Review*, 10(2):115–152, 1995.

[130] M. Wooldridge and N.R. Jennings, editors. *Intelligent Agents – Agent Theories, Architectures, and Languages*, volume 890 of *Lecture Notes in Computer Science (subseries LNAI)*. Springer-Verlag, 1995.

[131] M. Wooldridge, J.P. Müller, and M. Tambe, editors. *Intelligent Agents Volume II – Agent Theories, Architectures, and Languages*, volume 1037 of *Lecture Notes in Computer Science (subseries LNAI)*. Springer-Verlag, 1996.

[132] G.H. von Wright. *Norm and Action.* Routledge & Kegan Paul, London, 1963.

[133] G.H. von Wright. The logic of action: A sketch. In N. Rescher, editor, *The Logic of Decision and Action.* University of Pittsburgh Press, 1967.

[134] G.H. von Wright. On so-called practical inference. In J. Raz, editor, *Practical Reasoning*, chapter III, pages 46–62. Oxford University Press, 1978.

Samenvatting

It doesn't have to be so complicated all of the time.

Banana Yoshimoto, *'NP'*.

In één zin samengevat gaat dit proefschrift over formele technieken die gebruikt kunnen worden bij het modelleren van menselijk gedrag. Met behulp van zo'n modellering kan niet alleen precies gekeken worden waardoor bepaald gedrag veroorzaakt wordt, maar ook kan gewenst gedrag formeel beschreven worden. Dit formeel beschrijven, of formeel specificeren, kan belangrijk zijn om het gedrag van entiteiten die geacht worden zich min of meer menselijk te gedragen precies vast te leggen. Voorbeelden van dit soort entiteiten zijn robots als Archie (de man van staal), maar ook software die computerprogramma's een menselijk interface geeft en de levensechte karakters die optreden in een virtual reality omgeving. Entiteiten die een menselijk gedrag (dienen te) vertonen worden vaak aangeduid met de Engelse term *agent*, afkomstig uit het Latijn, hetgeen in essentie niets meer betekent dan 'handelende entiteit'. Aangezien een pakkende Nederlandse vertaling van deze term ontbreekt (het Van Dale Groot Woordenboek Engels-Nederlands suggereert 'handelend persoon' en het in de Nederlandse vertaling van de titel gebruikte 'actoren' is het ook niet helemaal) zullen we in de rest van deze samenvatting de Engelse termen 'agent' en 'agents' gebruiken, waarbij voor het gemak aangenomen wordt dat agents mannelijk zijn.

De laatste jaren wordt er zeer veel onderzoek gedaan naar agents, zowel op theoretisch als op praktisch niveau. Ondanks het feit dat er veel mensen bezig zijn met agents (of misschien juist wel doordat er zoveel mensen mee bezig zijn), bestaat er geen overeenstemming over wat nu precies onder een agent verstaan moet worden. Soms wordt een algorithme beschouwd als een agent, bij object-georiënteerd programmeren wordt een object wel eens gezien als een agent, terwijl agents vaak ook gezien worden als robots die zich als mensen gedragen. In dit proefschrift verstaan we onder een agent iedere entiteit die de mogelijkheid heeft bepaalde acties uit te voeren, de beschikking heeft over bepaalde informatie en redenen heeft om zich op een bepaalde manier te gedragen. Vaak zullen agents ook nog rationeel zijn in hun gedragingen, tot op zekere hoogte autonoom

zijn, en in staat zijn informatie te verwerven en die te gebruiken om hun gedrag aan te passen. De modellering van dit soort agents vormt het onderwerp van dit proefschrift.

Zoals reeds eerder opgemerkt zijn modelleertechnieken op tenminste twee manieren nuttig en bruikbaar. Enerzijds kunnen ze gebruikt worden om geconstateerd gedrag te analyseren, anderzijds zijn ze te gebruiken om gewenst gedrag vast te leggen. Beide aspecten zijn belangrijk bij de toepassing van formele technieken voor agents. Zo zijn er gevallen bekend waar met behulp van formele technieken de oorzaak is opgespoord die er voor zorgde dat bepaalde software agents zich afwijkend en ongewenst gedroegen. Daarnaast worden agents vaak toegepast in omgevingen waar het noodzakelijk is zekerheid omtrent hun gedrag te verkrijgen. Voorbeelden hiervan zijn toepassingen van agents in systemen die luchtverkeersleiders assisteren. Het gebruik van formele technieken in de specificatie van deze agents kan zorgen voor absolute zekerheid omtrent hun gedrag in alle mogelijke omstandigheden.

De formele systemen die in dit proefschrift gebruikt worden zijn gebaseerd op *modale logica's*. Oorspronkelijk zijn deze logica's door Leibniz voorgesteld om filosofische begrippen als 'noodzakelijkheid' en 'mogelijkheid' te representeren. Met name na de introductie van de *mogelijke-werelden semantiek* door Kripke worden modale logica's gebruikt om een breed scala van begrippen, zowel uit de filosofie als uit de informatica, te modelleren. In dit proefschrift worden bestaande modale logica's op een nieuwe wijze gecombineerd, en worden daarnaast verscheidene uitbreidingen van deze logica's voorgesteld.

In hoofdstuk 3 definiëren we het eerste, meest eenvoudige, formele systeem. In dit systeem is het mogelijk de kennis en vaardigheden van agents te modelleren, de resultaten van de acties die ze mogelijk uitvoeren, en het al dan niet hebben van de gelegenheid om een actie uit te voeren. Hierbij zien we kennis als ware informatie, die verder zo is dat een agent zich zowel bewust is van de dingen die hij weet als van de dingen die hij niet weet. Als een agent bijvoorbeeld weet dat 174306 even is maar niet weet of het deelbaar is door 417, dan is 174306 niet alleen inderdaad even maar weet de agent ook dat hij dit weet terwijl hij ook weet dat hij niet weet of het deelbaar is door 417. Alles wat veroorzaakt wordt door het uitvoeren van een actie wordt beschouwd als een deel van het resultaat van die actie. Het is bijvoorbeeld zo dat een resultaat van het verbranden van een brief is dat er rook ontstaat, maar ook dat er papier omgezet wordt in as. De combinatie van vaardigheid en gelegenheid bepaalt welke acties voor een agent praktisch (on)mogelijk zijn. De vaardigheden van een agent bepalen welke acties binnen zijn capaciteiten liggen; of de agent de gelegenheid heeft om die acties ook uit te voeren is afhankelijk van externe omstandigheden. Een voorbeeld dat het verschil en het verband tussen vaardigheid en gelegenheid duidelijk maakt betreft een leeuw in een dierentuin. Deze leeuw heeft zeer waarschijnlijk wel de vaardigheid een zebra te verscheuren, maar zal (zo is te hopen voor de zebra) niet de gelegenheid daarvoor krijgen. Om kennis, vaardigheden, gelegenheden

en resultaten te modelleren, combineren we *epistemische* logica, de modale logica van kennis, met *dynamische* logica, de modale logica van actie, waaraan een extra component toegevoegd is welke de vaardigheden van agents representeert.

In hoofdstuk 4 breiden we het basissysteem zo uit dat het mogelijk wordt acties te modelleren waarvan voor uitvoering nog niet vaststaat waar ze precies uit zullen bestaan. Dit soort acties wordt *niet-deterministisch* genoemd. Een standaard voorbeeld van een niet-deterministische actie is 'post de brief of verbrand hem'. Als alleen maar bekend is dat een agent deze actie gaat uitvoeren is nog niet duidelijk of de brief gepost of verbrand gaat worden. Hierbij is het van belang waardoor bepaald wordt hoe een niet-deterministische actie uitgevoerd dient te worden, en met name of de agent die de actie uitvoert enige invloed op deze keuze uit kan oefenen. In hoofdstuk 4 onderscheiden we twee soorten niet-deterministische acties die verschillen met betrekking tot degene die de keuze maakt: in het ene geval wordt de keuze door de agent gemaakt, in het andere geval wordt de keuze toegeschreven aan een externe omgeving waarop de agent geen invloed heeft. Voor een software interface agent zou deze externe omgeving een gebruiker kunnen zijn: het is mogelijk dat bepaalde acties die de agent dient uit te voeren vanuit het oogpunt van de agent niet-deterministisch zijn, waarbij de keuze bij de gebruiker ligt. Het blijkt dat standaard benaderingen die voorgesteld zijn voor het representeren van niet-determinisme niet bruikbaar zijn voor onze doeleinden; als gevolg hiervan zijn de benaderingen die wij voorstellen dan ook niet erg standaard.

In hoofdstuk 5 beschouwen we agents die zich bezighouden met allerlei aspecten van informatie en informatie-beheer. Deze agents kunnen bijvoorbeeld de elektronische post van een gebruiker beheren, of hem/haar begeleiden bij zoektochten op het internet. De informatie-beherende agents die wij beschouwen, beschikken naast kennis nog over zwakkere vormen van informatie. Ook hebben ze de mogelijkheid bepaalde acties uit te voeren die gericht zijn op het verwerven of overdragen van informatie. Zo kunnen de agents informatie verwerven door observaties en door het voor waar aannemen van bepaalde waarschijnlijke — maar net niet helemaal zekere — beweringen. Beweringen van dit soort worden *defaults* genoemd. Het bekendste voorbeeld van een default gaat over de vogel Tweety. Zolang er niets meer over Tweety bekend is dan dat het een vogel is, kan bij default aangenomen worden dat Tweety vliegt. Nieuwe informatie, bijvoorbeeld dat Tweety een pinguïn is, of een gebraden eend, kan ervoor zorgen dat de default-conclusie dat Tweety vliegt weer ingetrokken moet worden. Via communicatie kan informatie overgedragen worden aan andere agents. De betrouwbaarheid van informatie hangt af van de manier waarop die verkregen wordt. In het algemeen is het zo dat informatie afkomstig uit observaties als het meest betrouwbaar beschouwd wordt, dat informatie verkregen uit communicatie iets minder betrouwbaar is, en dat het aannemen bij default tot de minst betrouwbare informatie leidt. In een agent die zijn gebruiker begeleidt bij zoektochten op het internet zijn veel van deze aspecten van informatie-beheer zichtbaar.

Deze agent observeert het gedrag van de gebruiker, vult dat eventueel aan met informatie die verkregen is via communicatie met andere agents of is aangenomen bij default, en doet aan de hand daarvan bepaalde suggesties aan de gebruiker die hij begeleidt. Zo zou hij bij default kunnen concluderen dat een bepaald bericht op het internet, waarvan het bestaan hem door een andere agent is medegedeeld, interessant is voor de gebruiker, omdat de agent eerder geconstateerd heeft dat de gebruiker in het algemeen interesse heeft voor berichten van dit soort. Mocht de agent nu expliciet te horen krijgen dat de gebruiker niet geïnteresseerd is, dan zal hij zijn default-conclusie intrekken. Er zijn twee opmerkelijke aspecten zichtbaar in onze modellering van informatie-beherende agents. Het eerste aspect betreft een uitbreiding van de standaard mogelijke-werelden semantiek die ons toestaat gradaties in de betrouwbaarheid van informatie te onderscheiden. Daarnaast stellen we een uitbreiding voor van de standaard interpretatie van acties zoals die gebruikt wordt in dynamische logica, de modale logica van actie. In onze meer algemene interpretatie kunnen we op elegante wijze de acties modelleren die observaties, communicatie en het maken van default-aannamen representeren.

Hoofdstuk 6 bevat een modellering van de *drijfveren* of *motieven* van agents. Deze drijfveren maken duidelijk wat een agent beweegt om zich op een bepaalde manier te gedragen. In het algemeen stellen wij dat agents gedreven zijn om hun onvervulde wensen te vervullen, waarbij een wens een of ander primitief verlangen is. Aanhangers van Aristoteles kunnen bij dit primitief verlangen denken aan het najagen van kennis, terwijl die van Freud waarschijnlijk het Lust-prinzip voor ogen zal staan. Omdat agents rationeel zijn, zullen ze zich niet zonder meer tot doel stellen ieder onvervulde wens te vervullen maar zich beperken tot die onvervulde wensen die in principe vervulbaar zijn. Als een agent zich eenmaal een doel gesteld heeft, kan hij besluiten zich tot het ondernemen van een bepaalde actie te committeren, dat wil zeggen dat de agent met zichzelf afspreekt, of aan zichzelf belooft, dat hij de actie uit zal gaan voeren. Voorwaarde hierbij is dat de agent weet dat uitvoering van deze actie inderdaad tot vervulling van zijn doel zal leiden. Afspraken tot het ondernemen van bepaalde acties worden genoteerd in de agenda van de agent. Op ieder moment is een agent aan zichzelf verplicht de acties uit te voeren die in zijn agenda staan. Als nu op een bepaald moment een agent tot de ontdekking komt dat een actie waartoe hij zich gecommitteerd heeft niet langer meer uitvoerbaar is of geen enkel doel meer dient, kan de agent besluiten zijn aan zichzelf gedane belofte tot uitvoering van de actie te verbreken. Als gevolg hiervan worden eventueel nog resterende afspraken die op deze belofte betrekking hebben uit de agenda van de agent verwijderd.

In het laatste hoofdstuk vatten we de bijdragen van dit proefschrift nog eens kort samen, geven we een voorbeeld van een specificatie van een software agent, en beschouwen we mogelijkheden voor vervolgonderzoek.

Curriculum Vitae

Bernardus van Linder

19 juni 1968
Geboren te Gemert.

augustus 1980 – juli 1981
Ongedeeld VWO aan het Titus Brandsma Lyceum te Oss.

augustus 1981 – juli 1986
Ongedeeld Gymnasium aan het Gymnasium Camphusianum te Gorinchem.
Diploma Gymnasium behaald 29 mei 1986.

september 1986 – juni 1987
Studie Technische Bedrijfskunde aan de Universiteit Twente te Enschede.

september 1987 – juni 1992
Studie Informatica aan de Katholieke Universiteit Nijmegen.
Propaedeuse diploma behaald 26 augustus 1988.
Doctoraal diploma behaald 26 juni 1992 (cum laude).

juli 1992 – september 1993
Assistent in Opleiding aan de Vakgroep Informatica van de Vrije Universiteit te Amsterdam.

september 1993 – juli 1996
Assistent in Opleiding aan de Vakgroep Informatica van de Universiteit Utrecht.

Index

This page intentionally left blank

This page intentionally left blank

A hand held over a candle in angst fuelled bravado
a carbon trail scores a moist stretched palm
Trapped in the indecision of another fine menu
and you sit there and ask me to tell you the story so far
This is the story so far

Shuffling your memories dealing your doodles in margins
you scrawl out your poems across a beermat or two
and when you declare the point of grave creation
They turn round and ask you to tell them the story so far
This is the story so far

And you listen with a tear in you eye
to their hopes and betrayals and your only reply
is Slàinte Mhath

Marillion, *'Slàinte Mhath'*.

This page intentionally left blank