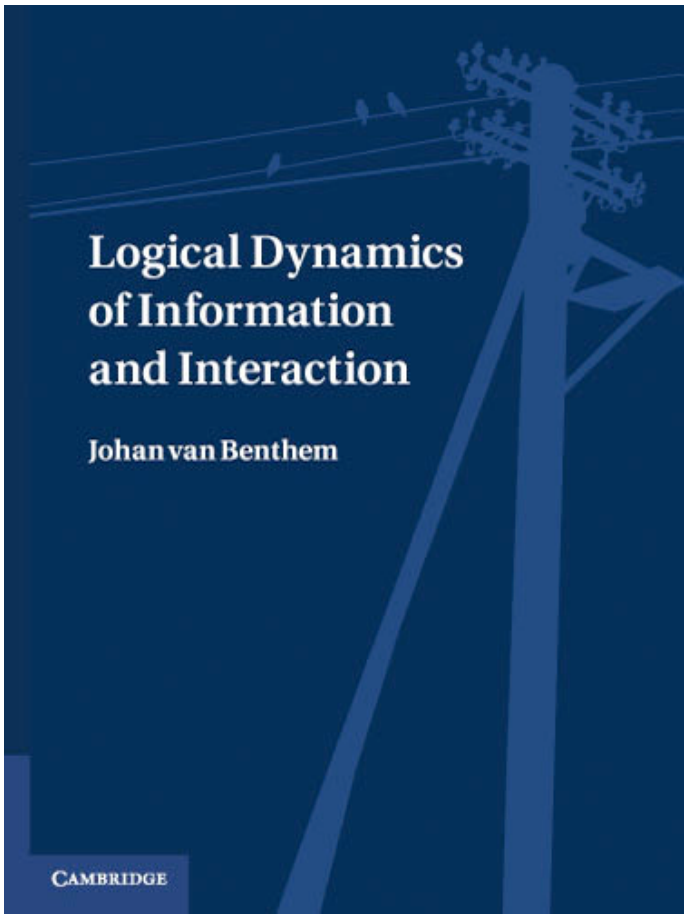
LOGICAL DYNAMICS OF INFORMATION AND INTERACTION

Johan van Benthem

Amsterdam & Stanford, http://staff.science.uva.nl/~johan/

Autumn 2011, Cambridge University Press



http://www.cambridge.org/us/academic/subjects/philosophy/logic/

logical-dynamics-information-and-interaction

TABLE OF CONTENTS

Preface

1 Logical dynamics, agency, and intelligent interaction

2 Epistemic logic and semantic information

3 Dynamic logic of public observation

4 Multi-agent dynamic-epistemic logic

5 Dynamics of inference and awareness

6 Questions and issue management

7 Soft information, correction, and belief change

8 An encounter with probability

9 Preference statics and dynamics

10 Decisions, actions, and games

11 Processes over time

12 Epistemic group structure and collective agency

13 Logical dynamics in philosophy

14 Computation as conversation

15 Rational dynamics in game theory

16 Meeting cognitive realities

17 Conclusion

Bibliography

To Arthur and Lucas

PREFACE

This book is about Logical Dynamics, a theme that first gripped me in the late 1980s. The idea had many sources, but what it amounted to was this: make actions of language use and inference first-class citizens of logical theory, instead of studying just their products or data, such as sentences or proofs. My program then became to explore the systematic repercussions of this 'dynamic turn'. It makes its first appearance in my book *Language in Action* (1991), where categorial grammars are linked to procedures of linguistic analysis using relational algebra – viewing natural language as a sort of cognitive programming language for transforming information. My next book *Exploring Logical Dynamics* (1996) continued with this perspective, linking it to modal logic and process theories in computer science: in particular, dynamic logic of programs. This added new themes like process invariances and definability, dynamic inference, and computational complexity of logics. In the meantime, my view of logical dynamics has evolved again. I now see it as a general theory of agents that produce, transform and convey information – and in all this, their social interaction should be understood just as much as their individual powers. Just think of this: asking a question and giving an answer is just as logical as drawing a conclusion on your own. And likewise, I would see argumentation with different players as a key notion of logic, with proof just a single-agent projection. This stance is a radical break with current habits, and I hope it will gradually grow on the reader, the way it did on me.

The book presents a unified account of the resulting agenda, in terms of *dynamic epistemic logic*, a framework developed around 2000 by several authors. Many of its originators are found in my references and acknowledgments, as are others who helped shape this book. In this setting, I develop a systematic way of describing actions and events that are crucial to agency, and show how it works uniformly for observation-based knowledge update, inference, questions, belief revision, and preference change, all the way up to complex social scenarios over time, such as games. In doing so, I am not claiming that this approach solves all problems of agency, or that logic is the sole guardian of intelligent interaction. Philosophy, computer science, probability theory, or game theory have important things to say as well. But I do claim that logic has a long-standing art of choosing abstraction levels that are sparse and yet revealing. The perspective offered here is simple, illuminating, and a useful tool to have in your arsenal when studying foundations of cognitive behaviour. Moreover, the logical view that we develop has a certain mathematical elegance that can be appreciated even when the grand perspective leaves you cold. And if that technical

appeal does not work either, I would already be happy if I could convey that the dynamic stance throws fresh light on many old things, helps us see new ones – and that it is fun!

This book is based on lectures and papers since 1999, many co-authored. Chapter 1 explains the program, Chapter 2 gives background in epistemic logic, and Chapters 3–12 develop the logical theory of agency, with a base line for readers who just wish to see the general picture, and extra topics for those who want more. Chapters 13–16, that can be read separately, explore repercussions of logical dynamics in other disciplines. Chapter 17 summarizes where we stand, and points at roads leading from here. In composing this story, I had to select, and the book does not cover every alley I have walked. Also, throughout, there are links to other areas of research, but I could not chart them all. Still, I would be happy if the viewpoints and techniques offered here would change received ideas about the scope of logic, and in particular, revitalize its interface with philosophy.

Acknowledgment First of all, I want to thank my co-authors on papers that helped shape this book: Cédric Dégrémont, Jan van Eijck, Jelle Gerbrandy, Patrick Girard, Tomohiro Hoshi, Daisuke Ikegami, Barteld Kooi, Fenrong Liu, Maricarmen Martinez, Stefan Minica, Siewert van Otterloo, Eric Pacuit, Olivier Roy, Darko Sarenac, and Fernando Velázquez Quesada. I also thank the students that I have interacted with on topics close to this book: Marco Aiello, Guillaume Aucher, Harald Bastiaanse, Boudewijn de Bruin, Nina Gierasimczuk, Wes Holliday, Thomas Icard, Lena Kurzen, Minghui Ma, Marc Pauly, Ben Rodenhäuser, Floris Roelofsen, Ji Ruan, Joshua Sack, Tomasz Sadzik, Merlijn Sevenster, Josh Snyder, Yanjing Wang, Audrey Yap, Junhua Yu, and Jonathan Zvesper. Also, many colleagues gave comments, from occasional to extensive, that improved the manuscript: Krzysztof Apt, Giacomo Bonanno, Davide Grossi, Andreas Herzig, Wiebe van der Hoek, Hans Kamp, Larry Moss, Bryan Renne, Gabriel Sandu, Sebastian Sequoiah-Grayson, Yoav Shoham, Sonja Smets, Rineke Verbrugge, and Tomoyuki Yamada. I also profited from the readers' reports solicited by Cambridge University Press, though my gratitude must necessarily remain de dicto. Finally, I thank Hans van Ditmarsch and especially Alexandru Baltag for years of contacts on dynamic epistemic logic and its many twists and turns.

Chapter 1

LOGICAL DYNAMICS, AGENCY, AND INTELLIGENT INTERACTION

1.1 Logical dynamics of information-driven agency

Human life is a history of millions of actions flowing along with a stream of information. We plan our trip to the hardware store, decide on marriage, rationalize our foolish behaviour last night, or prove an occasional theorem, all on the basis of what we know or believe. Moreover, this activity takes place in constant interaction with others, and it has been claimed that what makes humans so unique in the animal kingdom is not our physical strength, nor our powers of deduction, but rather our planning skills in social interaction – with the Mammoth hunt as an early example, and legal and political debate as a late manifestation. It is this intricate cognitive world that I take to be the domain of logic, as the study of the invariants underlying these informational processes. In particular, my program of *Logical Dynamics* (van Benthem 1991, 1996, 2001) calls for identification of a wide array of informational processes, and their explicit incorporation into logical theory, not as didactic background stories for the usual concepts and results, but as first-class citizens. One of the starting points in that program was a pervasive ambiguity in our language between *products* and *activities* or processes. "Dance" is an activity verb, but it also stands the product of the activity: a waltz or a mambo. "Argument" is a piece of a proof, but also an activity one can engage in, and so on. Logical systems as they stand are product-oriented, but Logical Dynamics says that both sides of the duality should be studied to get the complete picture. And this paradigm shift will send ripples all through our standard notions. For instance, natural language will now be, not a static description language for reality, but a dynamic programming language for changing cognitive states.

Recent trends have enriched the thrust of this action-oriented program. 'Rational agency' stresses the transition from the paradigm of proof and computation performed by a single agent (or none at all) to agents with abilities, goals and preferences plotting a meaningful course through life. This turn is also clear in computer science, which is no longer about lonely Turing Machines scribbling on tapes, but about complex intelligent communicating systems with goals and purposes. Another recent term, 'intelligent interaction', emphasizes what is perhaps the most striking feature here, the role of *others*. Cognitive powers show at their best in many-mind, rather than single-mind settings – just as physics only gets

interesting, not with single bodies searching for their Aristotelean natural place, but on the Newtonian view of many bodies influencing each other, from nearby and far.

1.2 The research program in a nutshell

What phenomena should logic study in order to carry out this ambitious program? I will first describe these tasks in general terms, and then go over them more leisurely with a sequence of examples. A useful point of entry here is the notion of *rationality*. Indeed, the classical view of humans as 'rational animals' seems to refer to our reasoning powers:

To be rational is to reason *intelligently*.

These powers are often construed narrowly as deductive skills, making mathematical proof the paradigm of rationality. This book has no such bias. Our daily skills in the common sense world are just as admirable, and much richer than proof, including further varieties of reasoning such as justification, explanation, or planning. But even this variety is not yet what I am after. As our later examples will show, the essence of a rational agent is the ability to use information from many sources, of which reasoning is only one. Equally crucial information for our daily tasks comes from, in particular, observation and communication. I will elaborate this theme later, but right now, I cannot improve on the admirable brevity of the Mohist logicians in China around 500 BC (Zhang & Liu 2007):

"Zhi: Wen, Shuo, Qin" 知 闻 说 亲 [1]

knowledge arises through hearing from others, inference, and observation.

Thus, while I would subscribe to the above feature of rationality, its logic should be based on a study of all basic informational processes as well as their interplay.

But there is more to the notion of rationality as I understand it:

To be rational is to act *intelligently*.

We process information for a purpose, and that purpose is usually not contemplation, but action. And once we think of action for a purpose, another broad feature of rationality comes to light. We do not live in a bleak universe of *information*. Everything we do, say, or perceive is coloured by a second broad system of what may be called *evaluation*, determining our preferences, goals, decisions, and actions. While this is often considered

[1] Somewhat anachronistically, I use modern simplified Chinese characters.

alien to logic, and closer to emotion and fashion, I would rather embrace it. Rational agents deal intelligently with both information and evaluation, and logic should get this straight.

Finally, there is one more crucial aspect to rational agency, informational and evaluational, that goes back to the roots of logic in Antiquity:

> *To be rational is to* interact *intelligently*.

Our powers unfold in communication, argumentation, or games: multi-agent activities over time. Thus, the rational quality of what we do resides also in how we interact with *others*: as rational as us, less, or more so. This, too, sets a broader task for logic, and we find links with new fields such as interactive epistemology, or agent studies in computer science.

I have now given rationality a very broad sense. If you object, I am happy to say instead that we are studying 'reasonable' agents, a term that includes all of the above. Still, there remains a sense in which mathematical deduction is crucial to the new research program. We want to describe our broader agenda of phenomena with *logical systems*, following the methods that have proven so successful in the classical foundational phase of the discipline. Thus, at a meta-level, in terms of modeling methodology, throughout this book, the reader will encounter systems obeying the same technical standards as before. And meta-mathematical results are as relevant here as they have always been. That, to me, is in fact where the unity of the field lies: not in a restricted agenda of 'consequence', or some particular minimal laws to hang on to, but in its methodology and modus operandi.

So much for grand aims. The following examples will illustrate what we are after, and each adds a detailed strand to our view of rational agency. We will then summarize the resulting research program, followed by a brief description of the actual contents of this book.

1.3 Entanglement of logical tasks: inference, update, and information flow

The Amsterdam Science Museum *NEMO* (http://www.nemo-amsterdam.nl/) organizes regular 'Kids' Lectures on Science', for some 60 children aged around 8 in a small amphitheatre. In February 2006, it was my turn to speak – and my first question was this:

> *The Restaurant* "In a restaurant, your Father has ordered Fish, your Mother ordered Vegetarian, and you have Meat. Out of the kitchen comes some new person carrying the three plates. What will happen?"

The children got excited, many little hands were raised, and one said: "He asks who has the Meat". "Sure enough", I said: "He asks, hears the answer, and puts the plate on the table. What happens next?" Children said: "He asks who has the Fish!" Then I asked once more what happens next? And now one could see the Light of Reason suddenly start shining in those little eyes. One girl shouted: "He does not ask!" Now, *that* is logic … After that, we played a long string of scenarios, including card games, Master Mind, Sudoku, and even card magic, and we discussed what best questions to ask and conclusions to draw.

Two logical tasks The Restaurant is about the simplest scenario of real information flow. And when the waiter puts that third plate without asking, you see a logical inference in action. The information in the two answers allows the waiter to infer (implicitly, in a flash of the 'mind's eye') where the third plate must go. This can be expressed as a logical form

$$A \text{ or } B \text{ or } C, \text{ not-}A, \text{ not-}B \Rightarrow C.$$

One can then tell the usual story about the power of valid inference in other settings. With this moral, the example goes back to Greek Antiquity. But the scenario is much richer. Let us look more closely: perhaps, appropriately, with the eyes of a child.

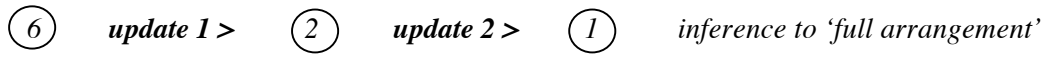
To me, the Restaurant cries out for a new look. There is a natural unity to the scenario. The waiter first obtains the right information by asking questions and understanding answers, acts of *communication* and perhaps *observation*, and once enough data have accumulated, he *infers* an explicit solution. Now on the traditional line, only the latter step of deductive elucidation is logic proper, while the former are at best pragmatics. But in my view, both informational processes are on a par, and both should be within the compass of logic. Asking a question and grasping an answer is just as logical as drawing an inference. And accordingly, logical systems should account for both of these, and perhaps others, as observation, communication, and inference occur entangled in most meaningful activities.

Information and computation And logic is up to this job, if we model the relevant actions appropriately. Here is how. To record the information changes in the Restaurant, a helpful metaphor is *computation*. During a conversation, information states of people – alone, and in groups – change over time, in a systematic way triggered by information-producing events. So we need a set of information states and transitions between them. And as soon as we do this, we will find some fundamental issues, even in the simplest scenarios.

Update of semantic information Consider the information flow in the Restaurant. The intuitive information states are sets of 'live options' at any stage, starting from the initial *6* ways of giving three plates to three people. There were two successive *update actions* on these states, triggered by the answers to the waiter's two questions. The first reduced the uncertainty from *6* to *2* options, and the second reduced it to *1*, i.e., just the actual situation. Here is a 'video' of how the answers for Meat and Fish would work in case the original order was ***FMV*** (fish for the first person, meat for the second, vegetarian for the third):

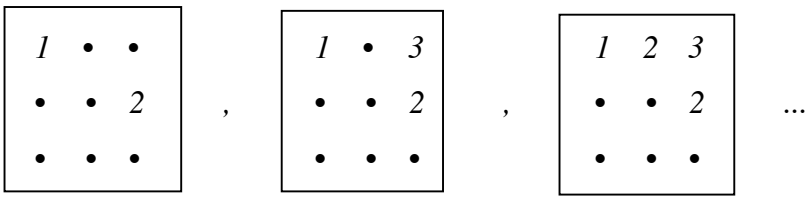
This is the common sense process of semantic update for the current *information range*, where new information is produced by events that rule out possibilities. In Chapters 2 and later, we will call this elimination scenario a case of 'hard information', and typical events producing it are public announcements in communication, or public observations.

Inference and syntactic information The first two updates have zoomed in on the actual situation. This explains why no third question is needed. But then we have a problem. What is the *point* of drawing a logical conclusion if it adds no further information? Here, the common explanation is that inferences 'unpack' information that we may have only implicitly. We have reached the true world, and now we want to spell it out in a useful sort of code. This is where inference kicks in, elucidating what the world looks like:



This sounds fine, but it makes sense only when we distinguish two different notions of information: one 'explicit', the other 'implicit' (van Benthem & Martinez 2008). Now, while there are elegant logics for semantic update of the latter, there is no consensus on how to model the explicit information produced by inference. Formats include syntactic accumulation of formulas, but more graphical ones also make sense. For instance, here is how propositional inferences drive stages in the solution of puzzles:

Example Take a simple 3x3 Sudoku diagram, produced by applying the two rules that 'each of the 9 positions must have a digit', but 'no digit occurs twice on a row or column':



Each successive diagram displays a bit more about the unique solution (one world) determined by the initial placement of the digits *1, 2*. Thus, explicit information is brought to light in logical inference in a process of what may be called deductive *elucidation*. Chapter 5 of this book will make a more systematic syntactic proposal for representing the dynamics of inference, that works in tandem with semantic update. For now, we just note that what happened at the Restaurant involves a basic issue in the philosophy of logic (cf. Chapter 13): capturing and integrating different notions of information.

Putting things together, the *dynamics* of various kinds of informational actions becomes a target for logical theory. But to make this work, we must, and will, also give an account of the underlying *statics*: the information states that the actions work over. As a first step toward this program, we have identified the first level of skills that rational agents have:

> *their powers of* inference*, and their powers of* observation*, resulting in information updates that change what they currently know*.

1.4 Information about others and public social dynamics

Another striking feature to the information flow in the Restaurant are the questions. Questions and answers typically involve more than one agent, and their dynamics is *social*, having to do also with what people come to know about each other. This higher-order knowledge about others is crucial to human communication and interaction in general.

Questions and answers Take just one simple "Yes/No" question followed by a correct answer, a ubiquitous building block of interaction. Consider the following dialogue:

> *Me*: "Is this Beihai Park?"
> *You*: "Yes."

This conveys facts about the current location. But much more is going on. By asking the question in a normal scenario (not, say, a competitive game), I indicate that I do not know the answer. And by asking you, I also indicate that I think you may know the answer, again

under normal circumstances. [2] Moreover, your answer does not just transfer bare facts to me. It also achieves that you know that I know, I know that you know that I know, and in the limit of such iterations, it achieves *common knowledge* of the relevant facts in the group consisting of you and me. This common knowledge is not a by-product of the fact transfer. It rather forms the basis of our mutual expectations about future behaviour. [3] Keeping track of higher-order information about others is crucial in many disciplines, from philosophy (interactive epistemology) and linguistics (communicative paradigms of meaning) to computer science (multi-agent systems) and cognitive psychology ('theory of mind'). [4] Indeed, the ability to move through an informational space keeping track of what other participants know and do not know, including the crucial ability to switch and view things from other people's perspectives, seems characteristic of human intelligence.

So, logical activity is interactive, and its theory should reflect this. Some colleagues find this alarming, as social aspects are reminiscent of gossip, status, and Sartre's "Hell is the Others". The best way of dispelling such fears may be a concrete example. Here is one, using a card game, a useful normal form for studying information flow in logical terms. It is like the Restaurant in some ways, but with a further layer of higher-order knowledge.

The Cards (van Ditmarsch 2000) Three cards 'red', 'white', 'blue' are distributed over three players: *1, 2, 3*, who get one each. Each player sees her own card, but not the others. The real distribution over *1, 2, 3* is *red, white, blue*. Now a conversation takes place (this actually happened during the *NEMO* children session, on stage with thee volunteers):

2 asks 1	*"Do you have the blue card?"*
1 answers truthfully	*"No"*.

Who knows what then, assuming the question is sincere? Here is the effect in words:

> "Assuming the question is sincere, *2* indicates that she does not know the answer, and so she cannot have the blue card. This tells *1* at once what the deal was. But

[2] All such presuppositions are off in a classroom with a teacher questioning students. The logics we develop in this book can deal with a wide variety of such informational scenarios.

[3] If I find your pin code and bank account number, I may empty your account – if I know that you do not know that I know all this. But if I know that you know that I know, I will not. Crime is triggered by fine iterated epistemic distinctions: that is why it usually takes experts.

[4] Cf. Hendricks 2005, Verbrugge 2009, van Rooij 2005, and many other sources.

> *3* did not learn, since he already knew that *2* does not have blue. When *1* says she does not have blue, this now tells *2* the deal. *3* still does not know the deal; but since he can perform the reasoning just given, he knows that the others know it."

We humans go through this sort of reasoning in many settings, with different knowledge for different agents. In Chapters 2, 3, we will analyze this information flow in detail.

These scenarios can be much more complex. Real games of 'who is the first to know' arise by restricting possible questions and answers, and we will consider game logics later on. Also, announcements raise the issue of the reliability of the speaker, as in logic puzzles with meetings of Liars and Truth-Tellers. Our systems will also be able to deal with these in a systematic way, though separating one agent type from another is often a subtle manner of design. Logic of communication is not easy, but it is about well-defined issues.

Thus, we have a second major aspect of rational agents in place as a challenge to logic:

> *their* social powers *of mutual knowledge and communication*.

Actually, these powers involve more than pure information flow. Questions clearly have other uses than just conveying information: they define *issues* that give a purpose to a conversation or scientific investigation. This dynamics, too, can be studied per se, and Chapter 6 will show how to deal with 'issue management' within our general framework.

1.5 Partial observation and differential information

The social setting suggests a much broader agenda for logical analysis. Clearly, public announcement as we saw in the Restaurant or with the Cards is just one way of creating new information. The reality in many games, and most social situations, is that information flows differentially, with partial observation by agents. When I draw a card from the stack, I see which card I am getting. You do not, though you may know it is one of a certain set: getting *some* information. When you take a peek at my card, you learn something by cheating, degrading my knowledge of the current state of the game into mere belief. When you whisper in your neighbour's ear during my talk, this is a public announcement in a subgroup – where I and others need not catch what you are saying, and I may not even notice that any information is being passed at all.

Modeling such information flow is much more complicated than public announcement, and goes beyond existing logical systems. The first satisfactory proposals were made only in the late 1990s, as we shall see in Chapter 4. By now, we can model information flow in

parlour games like "Clue", that have an intricate system of public and private moves. All this occurs in natural realities all around us, such as *electronic communication*:

> "I send you an email, with the message *'P'*: a public announcement in the group {*you, me*}, You reply with a message *'Q'* with a *cc* to others: a public announcement to a larger group. I respond with *'R'* with a 'reply-to-all' plus a *bcc* to some further agents."

In the third round, we have a partly hidden act again: my *bcc* made an announcement to some agents, while others do not know that these were included. The information flow in this quite common episode is not simple. After a few rounds of *bcc* messages to different groups, it becomes very hard to keep track of who is supposed to know what. And that makes sense: differential information flow is complex, and so is understanding social life.

There are intriguing thresholds here. Using *bcc* is not misleading to agents who know that it is a possible event in the system. A further step is *cheating*. But even judicious lies seem a crucial skill in civilised life. Our angelic children are not yet capable of that, but rational agents at full capacity can handle mixtures of lies and truths with elegance and ease.

Thus, we have a further twist to our account of the powers of rational agents:

> different observational access *and processing differential information flow*.

This may seem mere engineering. Who cares about the sordid realities of cheating, lying, and social manoevering? Well, differential information is a great good: we do not tell everyone everything, and this keeps things civilized and efficient. Indeed, most successful human activity is social, from hunting cave bears to mathematics. And a crucial feature of social life is organization, including new procedures for information flow. Even some philosophy departments now do exams on Skype, calling for new secret voting procedures on a public channel. What is truly amazing is how this fascinating informational reality has been such a low priority of mainstream logicians and epistemologists for so long.

1.6 Epistemic shocks: self-correction and belief revision

So far, we considered information flow and knowledge. It is time for a next step. Agents who correctly record information from their observations, and industriously draw correct conclusions from their evidence, may be rational in some Olympian sense. But they are still cold-blooded recording devices. But knowledge is scarce, and rationality does not reside in always being cautious, and continual correctness. Its peak moments occur with warm-blooded agents, who are opinionated, make mistakes, but who subsequently *correct*

themselves. [5] Thus, rationality is about the dynamics of being wrong just as much as about that of being right: through belief revision, i.e., *learning* by giving up old beliefs. [6] Or maybe better, rationality is about a balance between two abilities: jumping to conclusions, and subsequent correction if the jump was over-ambitious.

Here, events become more delicate than with information flow through observation. Our knowledge can never be falsified by true new information, but beliefs can, when we learn new facts contradicting what we thought most plausible so far. Feeling that an earthquake is hitting the Stanford campus, I no longer believe that a short nocturnal bike ride will get me home in 10 minutes. There is much for logic to keep straight in this area. For instance, the following nasty scenario has been discussed by computer scientists, philosophers, and economists in the 1990s. Even true beliefs can be sabotaged through true information:

> *Misleading with the Truth*
>
> You know that you have finished *3rd, 2nd,* or *1st* in the election, and you find lower outcomes more plausible. You also know that being *2nd* makes your bargaining chances for getting some high office small ('dangerous heavy-weight'). In fact you were *1st*. I know this, but only say (truly) that you are not *3d*: and you become unhappy. Why?

Initially, you find being *3rd* the most plausible outcome, and may believe you will get high office by way of compensation. So, this is a true belief of yours, but for the wrong reason. Now you learn the true fact that you are not *3rd*, and being *2nd* becomes the most plausible world. But then, you now believe, falsely, that you will not have any high office – something you would not believe if you knew that you won the election.

Our logics of belief revision in Chapter 7 of this book can deal with such scenarios. They even include others with a softer touch, where incoming *soft information* merely makes certain worlds less or more plausible, without ever removing any world entirely from consideration. In this same line, Chapter 8 will show how dynamic logics can also incorporate *probability*, another major approach to beliefs of various strengths.

---

[5] Compare a lecture with a mathematician writing a proof on a blackboard to a research colloquium with people guessing, spotting problems, and then making brilliant recoveries…

[6] In a concrete setting, revision arises in *conversation*. People contradict each other, and then something more spectacular is needed than update. Maybe one of them was wrong, maybe they all were, and they have to adjust. Modeling this involves a further distinction between information coming from some source, and agents' various attitudes and responses to it.

Thus, in addition to the earlier update that accumulates knowledge, we have identified another, more complex, but equally important feature of rational agents:

> *their capacity for hypothesizing, being wrong, and then* correcting themselves.

In many settings, these capacities seems the more crucial and admirable human ability. A perfectly healthy body is great, but lifeless, and the key to our biological performance is our immune system responding to cuts, bruises, and diseases. Likewise, I would say that flexibility in beliefs is essential: and logic is all about the immune system of the mind.

1.7 Planning for the longer term

So far, we have mostly discussed single moves that rational agents make in response to incoming information, whether knowledge update or belief revision. But in reality, these single steps make sense only as part of longer processes through time. Even the Restaurant involved a conversation, that is, a sequence of steps, each responding to earlier ones, and directed toward some goal, and the same is true for games and social activities in general. There is relevant structure at this level, too, and as usual, it is high-lighted by well-known puzzles. Here is an evergreen:

> *The Muddy Children* (Fagin et al. 1995):
>
> After having played outside, two of three children have got mud on their foreheads. They can only see the others, so they do not know their own status. [7]Now their Father comes along and says: “At least one of you is dirty”. He then asks: “Does anyone know if he is dirty?" Children answer truthfully, and this is repeated round by round.
>
> As questions and answers repeat, what will happen?

One might think that nothing happens, since the father just tells the children something they already know – the way parents tend to do – viz. that there is at least one dirty child. But in reality, he does achieve something significant, making this fact into *common knowledge*. Compare the difference between every colleague knowing that your partner is unfaithful: no doubt unpleasant, but maybe still manageable, with this fact being common knowledge, including everyone knowing that the others know, etcetera: the shame at department meetings becomes unbearable. Keeping this in mind, here is what happens:

---

[7] This observational access is the inverse of our earlier card games, but formally very similar.

> Nobody knows in the first round. But in the next round, each muddy child reasons like this: "If I were clean, the one dirty child I see would see only clean children, and so she would know that she was dirty. But she did not. So I must be dirty, too!"

Note that this scenario is about what happens in the long run: with more children, common knowledge of the muddy children arises after more rounds of ignorance announcement, after which, in the next step, the clean children will know that they are clean. There is a formal structure to this. The instruction to the children looks like a computer program:

> REPEAT (IF you don't know your status THEN say you don't know ELSE say you do).

This is no coincidence. Conversation involves *plans*, and plans have a control structure for actions also found in computer programs: choice, sequential composition, and iteration of actions. The muddy children even have *parallel composition* of actions, since they answer simultaneously. Thus, actions may be composed and structured to achieve long-term effects – and this, too, will be an aspect of our logics. But for the moment, we note this:

> *Rational agency involves planning in* longer-term scenarios*, and its quality also lies in the ways that agents compose their individual actions into larger wholes*.

We will study long-term perspectives on agent interaction in Chapter 11, with connections to temporal logics of branching time as the Grand Stage where human activity takes place: as in Jorge Luis Borges' famous story 'The Garden of Forking Paths'.

1.8 Preferences, evaluation, and goals

Now we move to another phenomenon, that is crucial to understanding the driving force of much informational behaviour as discussed so far. Just answering 'a simple question' is rare. Behind every question, there lies a *why*-question: why does this person say this, what does she want, and what sort of scenario am I entering? Pure informational activities are rare, and they tend to live in an ether of preferences, and more generally, *evaluation* of situations and actions. This is not just greed or emotion. 'Making sense' of an interaction involves meaning and information, but also getting clear on the goals of everyone involved. This brings in another level of agent structure: crucially,

> *logic of rational agency involves* preferences *between situations and actions and agents'* goals, *usually aligned with these preferences*.

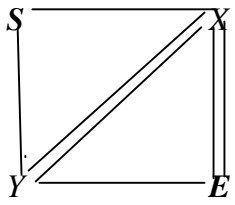
Preferences determine actions, and knowing your preferences helps me make predictions about what will happen. [8] It is hard to separate information from evaluation, and this may reflect some deep evolutionary entanglement of our cognitive and emotional brain systems.

While preference has been studied extensively in decision theory and game theory, it has been more marginal in logic. In Chapter 9 of this book, we incorporate preference logic, and show how it fits well with logical analysis of knowledge update and belief revision. Indeed, it might be said that this provides the explanatory dynamics in the physicists' sense behind the 'kinematics' of knowledge and belief that we have emphasized so far. [9]

1.9 Games, strategies, and intelligent interaction

Temporal perspective and preference combine in the next crucial feature of rational agents that we noted earlier, their responding to others and mutually influencing them. Even a simple conversation involves choosing assertions depending on what others say. This interactive aspect means that dynamic logics must eventually come to turn with *games*:

True interaction and games To sample the spirit of interaction, consider the following game played between a Student and a Teacher. The Student is located at position *S* in the following diagram, but wants to reach the position of escape *E* below, whereas the Teacher wants to prevent him from getting there. Each line segment is a path that can be traveled. At each round of the game, the Teacher cuts one connection, anywhere in the diagram, while the Student can, and must travel one link still open to him at his current position:



If Teacher is greedy, and starts by cutting a link *S–X* or *S–Y* right in front of the Student, then Student can reach the escape *E*. However, Teacher does have a *winning strategy* for preventing the Student from reaching *E*, by doing something else:

---

[8] If you see this is daily life, not science, think of how a referee judges your paper on its interest rather than truth, where 'interest' depends on the preferences of a scientific community.

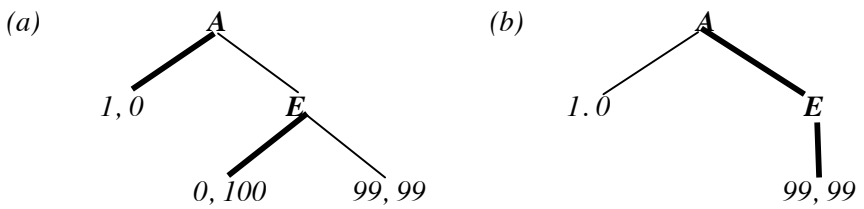
[9] I take this analogy seriously. *Rationality* (cf. Chapters 10, 15) will link preference, belief and action in ways reminiscent of Newton's Laws for the dynamics of moving bodies using force, mass and acceleration. We mix theoretical and observational terms, and then base explanations on them.

first cutting one line between *X* and *E*, and then letting his further cutting be guided in a straightforward manner by where Student goes subsequently.

Here *strategies* for players are rules telling them what to do in every eventuality. Solving games like this can be complex, emphasizing the non-trivial nature of interaction. [10]

Digression: learning Formal Learning Theory (Kelly 1996) concentrates on single-agent settings where a student forms hypotheses on the basis of some input stream of evidence: there is a Student, but no Teacher, unless we think of Nature as a disillusioned teacher doing a minimum of presentation without adjustment. But the realities of teaching and learning are social, with Students and Teachers responding to each other – and learning is a social process. We even learn at two levels: 'knowledge that', and know-how or skills.

Logic and game theory These multi-agent scenarios are close to game theory (Osborne & Rubinstein 1994 is an excellent introduction whose style also speaks to logicians) where information, evaluation, and strategic interaction are entangled. For a start, *Zermelo's Theorem* says that extensive two-player games of finite depth with perfect information and zero-sum outcomes are *determined*: that is, one of the players has a winning strategy. In our teaching game, this explains why Student or Teacher has a winning strategy. [11] Real game theory arises when players have preferences and evaluate outcomes. The reasoning extending Zermelo's is *Backward Induction*. Starting from values on leaves, nodes get evaluated up the tree, representing players' intermediate beliefs as to expected outcomes, given that both are acting 'rationally'. Here is an example, with nodes indicating the turns of two players ***A***, ***E***, while branching indicates different available moves. At end nodes, values are indicated for the players, in the following order ('value for ***A***', 'value for ***E***'):



[10] Rohde 2005 shows that solving 'sabotage graph games' is *Pspace*-complete, a high degree of complexity. The reader will get a better feel for the complexity of interaction by considering a variant. This time, the Teacher wants to force the Student to *end up in E* without any possibility of escape. Who of the two has the winning strategy this time, in the same graph?

[11] For details of this and the next examples, cf. Chapter 10.

The thick black lines in Tree (a) indicate the backward induction moves of 'rational' players who choose those actions whose outcomes they believe to be best for themselves. Interestingly, this is an equilibrium with a socially undesirable outcome, as *(1, 0)* makes both players worse off than *(99, 99)*. Thus, we need to reassess the assumptions behind the usual solution procedures for games. Dynamic logics of communication help here, with new takes. Think of *promises* that change a game by announcement of intentions. ***E*** might promise that she will not go left, changing game (a) to game (b) – and the new equilibrium *(99, 99)* results, making both players better off by restricting the freedom of one. [12]

Finally, games are not just an analytical tool. They are also a ubiquitous human activity across cultures, serving needs from gentle elegant wastes of time to training crucial skills:

> *a full logical understanding of rational agency and intelligent interaction*
> *requires a logical study of* games*, as a crucial model for human behaviour*.

This theme is mostly the subject of van Benthem, to appearA, but Chapters 10, 11, 15 take it up in some detail, including games with partial observation and imperfect information.

1.10 Groups, social structure, and collective agency

Single agents need not just interact on their own: typically, they also form *groups* and other collective agents, whose behaviour does not reduce to that of individual members. For instance, groups of players in games can form coalitions, and social choice theory is about groups creating group preferences on the basis of individual preferences of their members. We saw some of this in the notion of common knowledge, which is about the degree of being informed inside a group. But there are more themes that concern group agency, and in Chapter 12 we will show how our dynamic logics interface with group behaviour, and even may help provide a 'micro-theory' of information-based rational deliberation.

1.11 The program of Logical Dynamics in a nutshell

Rational agency involves information flow with many entangled activities: inference, observation, communication, and evaluation, all over time. Logical Dynamics makes all of these first-class citizens, and says that logical theory should treat them on a par. The resulting dynamic logics add subtlety and scope to classical systems, going beyond agent-

[12] Van Benthem 2007F proposes alternatives to Backward Induction in history-oriented games, where players remind themselves of the *legitimate rights of others*, or of past favours received.

free proof and computation. [13] Thus, we get new interdisciplinary links beyond old friends of mathematics, philosophy, [14] and linguistics, including computer science and economics. While reclaiming a broader agenda, with formal systems a means but not the end, logic also becomes a central part of academic life, overflowing the usual disciplinary boundaries.

A historical pedigree Is this new-fangled tinkering with the core values of logic? I do not think so. The ideas put forward here are ancient. We already mentioned broader Chinese views from Mohist logic. Likewise, traditional Indian logic stressed three ways of getting information. The easiest route is to observe, when that is possible. The next method is inference, in case observation is impossible or dangerous – the example being a coiled object in a room where we cannot see if it is a piece of rope, or a cobra. And if these two methods fail, we can resort to communication, and ask an expert. And also in the Western tradition, the social interactive aspect of information flow was there from the start. While many see Euclid's *Elements* as the paradigm for logic, with its crystalline mathematical proofs and eternal truths, the true origin may be closer to Plato's *Dialogues*, an argumentative practice. It has been claimed that logic arose out of legal, philosophical, and political debate in all its three main traditions. [15] And this multi-agent interactive view has emerged anew in modern times. A beautiful case are the *dialogue games* of Lorenzen 1955, that explained logical validity pragmatically in terms of a winning strategy for a proponent arguing a conclusion against an opponent granting the premises. Similar views occur in Hintikka 1973, another pioneer of games and informational activities inside logic.

But in the end, I see no opposition between Platonic and Euclidean images. This book is about a broad range of logical activities, but still pursued by mathematical means. Logical dynamics has no quarrel with classical standards of explicitness and precision.

The promised land? The area staked out here may be the logician's Promised Land, but it is hardly virgin territory. Like Canaan in the Old Testament, it is densely settled by other nations, such as philosophers, computer scientists, or economists, worshipping other gods

---

[13] Moreover, we find a side benefit to the Dynamic Turn. Chapter 13 shows how to replace 'non-standard logics', that have sprung up in droves in recent years, by perfectly classical systems, once we identify the right information-changing events, and make them an explicit part of the logic.

[14] In particular, I see strong connections with (social) *epistemology*, cf. Goldman 1999.

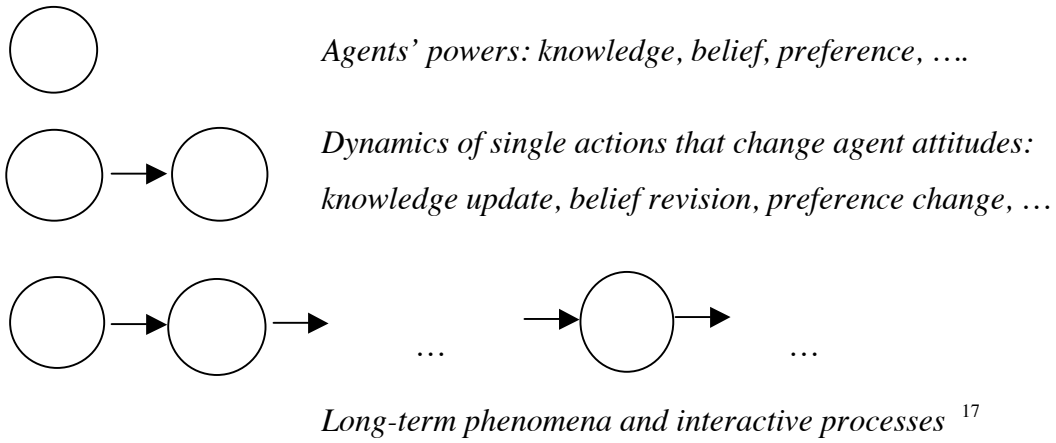
[15] The Mohists in early China discuss the Law of Non-Contradiction as a principle of conversation: 'resolve contradictions with others', 'avoid contradicting yourself', Zhang & Liu 2007.

such as probability or game theory. Can we just dispossess them? Indeed we must not. Rational agency is a deep subject, calling for all the help we can. I think logic has fresh insights to offer, beyond what is already there. But it has no favoured position: polytheism is a civilized idea. I do predict new fruitful liaisons between logic and its neighbours. [16]

A bridge too far? On top of the logical micro-structure and discrete temporal processes that we will study, there is emergent statistical behaviour of large groups over a long time. That is the realm of probability, evolutionary games and *dynamical systems*. The intriguing interface of dynamic logics and dynamical systems is beyond the scope of this book.

I.12 The chapters explained

Some students find it helpful to think of the whole program here pictorially, in stages:



[17]

Our chapters follow these lines, high-lighting activities and powers of agents. We start with semantic information and update, taking knowledge in the relaxed sense of what is true according to the agent's hard information. Our tools are epistemic logic over possible worlds model in a concrete sense (Chapter 2), its dynamified version of public announcement logic (*PAL*; Chapter 3) for communication and observation, and the more sophisticated dynamic-epistemic logic (*DEL*) of Chapter 4 with private scenarios and many agents. These systems are our paradigm for analysis of definable changes in current models of the agents' information. This methodology is applied in Chapter 5 to deal with inference, and other actions turning implicit into explicit knowledge. Chapter 6 shows how this also works for questions and issue management. Next, we turn to belief revision in

[16] For a wonderful sample of what can be learnt by combining disciplines, cf. Shoham & Leyton-Brown 2007 on multi-agent systems in terms of computer science, game theory, and logic.

[17] One might add a second dimension of *group size*, but it would overload the picture.

Chapter 7, using changes in plausibility orders to model learning systematically. Chapter 8 is a digression, showing how similar ideas work for probabilistic update, with a richer quantitative view of learning mechanisms. Then we move to agents' evaluation of worlds and actions, and show how our techniques for plausibility change also apply to preferences and ways of changing them, providing a unified account of information, belief, and goals (Chapter 9). Next, moving beyond one-step dynamics, Chapter 10 is about longer-term multi-agent interaction, with a special emphasis on games. A still wider perspective is that of Chapter 11, with embeddings of our dynamic logics in epistemic temporal logics of branching time, and logics of protocols that add 'procedural information' to our study of agency. Completing the development of our theory, Chapter 12 considers groups as new logical agents, showing how the earlier systems lift to this setting, including new phenomena such as belief merge, as well as links with social choice theory.

Advice to the reader These chapters are all arranged more or less as follows. First comes the motivation, then the basic system, then its core theory, followed by a conclusion explaining how one more building block of our logic of agency has been put in place. Readers could opt out at this stage, moving on to the next topic in the chapter sequence. What follows in a chapter is usually a logician's pleasure garden with further technical themes, open problems, and a brief view of key literature. Our open problems are both technical and conceptual – and non-logicians, too, may find some of them worthwhile.

The remaining Chapters 13–16 show how the logical dynamics developed here applies to a range of disciplines. Chapter 13 is concerned with philosophy, putting many old issues in a fresh light, and in that same light, adding new themes. Chapters 14–16 extend the interface to computer science, game theory, and cognitive science. These chapters can be read independently: there is no sequence, and they are different in style and level of technicality. Finally, Chapter 17 states our main conclusions and recommendations.

Chapter 2 EPISTEMIC LOGIC AND SEMANTIC INFORMATION

Our first topic in the study of agency is the intuitive notion of information as a *semantic range of possibilities*, widespread in science, but also a common sense view. For this purpose, we use epistemic logic, proposed originally for analyzing the philosophical notion of knowledge (Hintikka 1962). While the latter use remains controversial, we will take the system in a neutral manner as a logic of semantic information. More precisely, the *hard information* that an agent currently has is a set of possible worlds, and what it 'knows' is that which is true in all worlds of that range (van Benthem 2005C). This picture makes knowledge a standard universal modality. We present some basics in this chapter, stressing points of method that will recur. We refer to Blackburn, de Rijke & Venema 2000, Blackburn, van Benthem & Wolter, eds., 2006 for all details in what follows. [18] We add a few special themes and open problems, setting a pattern that will return in later chapters.

The topics to come reflect different aspects of a logical system. Its language and semantics provide a way of describing situations, evaluating formulas for truth or falsity, engaging in communication, and the like. This leads to issues of definition and *expressive power*. Next there is the *calculus of valid reasoning*, often with completeness theorems tying this to the semantics. In this book, this theme will be less dominant. To me, modal logics are not primarily about inferential life-styles like *K, KD45,* or *S5*, but about describing agency. Finally, there is *computational complexity*. We want to strike a balance between expressive power and complexity of logical tasks (cf. van Benthem & Blackburn 2006). This comes out well in a procedural perspective on meaning and proof in the form of 'logic games' – an interesting case of interactive methods entering logic. Our presentation is not a textbook for epistemic logic, [19] but a showcase of a standard logical system that we will 'dynamify' in this book – not to replace it with something else, but to make it do even further things.

At this point, readers who know their epistemic logic might skip to the next chapters.

2.1 The basic language

We start with the simplest epistemic language, describing knowledge of individual agents,

[18] Epistemic logic is flourishing today in computer science, game theory, and other areas beyond its original habitat: cf. Fagin et al. 1995, van der Hoek & Meijer 1995, van der Hoek & Pauly 2007.

[19] A new textbook presentation of modal logic in the current spirit is van Benthem, to appearB.

taken from some set *I* of agents that are relevant to the application at hand. But let us first illustrate what sort of situation this language is typically supposed to describe.

Example Questions and Answers.
A stranger approaches you in Beijing, and asks

Q "Is this Beihai Park?"

As a well-informed and helpful Chinese citizen, you answer truly

A "Yes."

As we noted in Chapter 1, this involves a mix of factual information and higher-order information about information of others. In particular, the answer produces *common knowledge* between ***Q*** and ***A*** of the relevant topographical fact. ■

It is worth having a logic that gets clear on these matters, and epistemic logic, first created to describe cogitation by lonesome thinkers, is just right for this interactive purpose. Thus, we can deal in a natural manner with the social aspects of agency that we noted before.

Definition Basic epistemic language.
The *basic epistemic language EL* has a standard propositional base with proposition letters and Boolean operators ('not', 'and', 'or', 'if then'), plus modal operators $K_i\varphi$ ('*i knows that* φ'), for each agent *i* in some set *I* that we fix in specific applications, and also $C_G\varphi$: 'φ is common knowledge in the group *G*' (with $G \subseteq I$). The inductive syntax rule is as follows, where the '*p*' stands for any choice of atoms from some set of proposition letters:

$$p \mid \neg\varphi \mid \varphi \wedge \psi \mid K_i\varphi \mid C_G\varphi$$

We write $\langle i \rangle \varphi$ for the existential modality $\neg K_i \neg \varphi$ saying intuitively that 'agent *i* considers φ possible'. The existential dual modality of $C_G\varphi$ is written as $\langle C_G \rangle \varphi$.[20] [21] ■

Through its agent-indexed modalities, this basic language describes the preceding scenario.

Example Questions and Answers, continued.
Let ***Q*** ask a factual question "φ?", to which ***A*** answers truly: "Yes". The presupposition for a normal truthful answer is that ***A*** knows that φ: written as $K_A\varphi$. The question itself, if it

[20] Individual modalities are first-order universal quantifiers, but common knowledge is a higher-order notion. In this initial phase of our presentation, this technical difference will not matter.

[21] We will also use other connectives $\vee, \rightarrow, \leftrightarrow$ as convenient, defined in the usual way.

is a normal co-operative one, conveys at least the presuppositions (i) $\neg K_Q\varphi \wedge \neg K_Q\neg\varphi$ (***Q*** does not know if φ) and (ii) $\langle Q\rangle(K_A\varphi \vee K_A\neg\varphi)$ (***Q*** thinks that ***A*** may know the answer). After the whole two-step communication episode, φ is known to both agents: $K_A\varphi \wedge K_Q\varphi$, while they also know this about each other: $K_QK_A\varphi \wedge K_AK_Q\varphi$, up to any depth of iteration. Indeed, they achieve the limit notion of common knowledge, written as $C_{\{Q,A\}}\varphi$. ■

Common knowledge is a group notion whose importance has been recognized in many areas: from philosophy to game theory, computer science, linguistics, and psychology.

2.2 Models and semantics

The preceding assertions only make precise sense when backed up by a formal semantics. Here is the formal version of the earlier intuitive idea of information as range.

Definition Epistemic models.

Models ***M*** for the language are triples $(W, \{\rightarrow_i \mid i\in I\}, V)$, where W is a set of *worlds*, the $\rightarrow_i$ are binary *accessibility relations* between worlds, and V is a *valuation* assigning truth values to proposition letters at worlds. In what follows, our primary semantic objects are *pointed models* $(\boldsymbol{M}, s)$ where s is the actual world representing the true state of affairs. ■

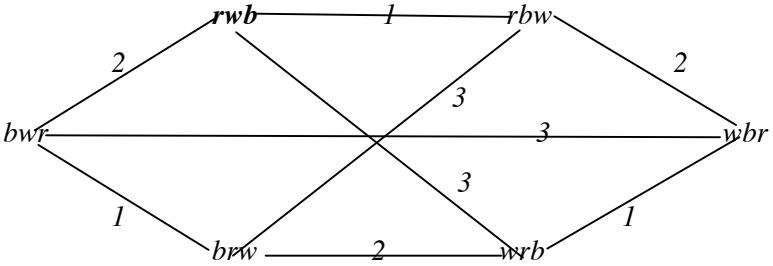
These models stand for collective information states of a group of agents. We impose no general conditions on the accessibility relations, such as reflexivity or transitivity – leaving a 'degree of freedom' for a modeler using the system. Many of our examples work well with equivalence relations (reflexive, symmetric, and transitive) – and such special settings may help the reader. But in some chapters, we will need arbitrary accessibility relations allowing end-points without successors, to leave room for false or misleading information.

Important notes on notation In some contexts, the difference between the semantic accessibility symbol $\rightarrow_i$ and a syntactic implication $\rightarrow$ is hard to see. To avoid confusion, we will often also use the notation $\sim_i$ for epistemic accessibility. The latter usually stands for an equivalence relation, while the arrow has complete generality. Also, we mostly write $\boldsymbol{M}, s$ for pointed models, using the brackets $(\boldsymbol{M}, s)$ only to remove ambiguity.

Example Setting up realistic epistemic models.

We will mainly use simple models to make our points. Even so, a real feel for the sweep of the approach only comes from the 'art of modeling' for real scenarios. Doing so also dispels delusions of grandeur about possible worlds. Consider this game from Chapter 1.

Three cards 'red', 'white', 'blue' were given to three players: *1, 2, 3*, one each. Each player can see her own card, but not that of the others. The real distribution over the players *1, 2, 3* is *red, white, blue* (written as ***rwb***). Here is the resulting information state:



This pictures the 6 relevant states of the world (the hands, or distributions of the cards), with the appropriate accessibilities (equivalence relations in this case) pictured by the uncertainty lines between hands. E.g., the single *1*-line between *rwb* and *rbw* indicates that player *1* cannot distinguish these situations as candidates for the real world, while *2* and *3* can (they have different cards in them). Thus, the diagram says the following. Though they are in ***rwb*** (as an outside observer can see), no player knows this. Of course, the game itself is a dynamic process yielding further information, as we will see in Chapter 3. ■

Over epistemic models, that may often be pictured concretely as a information diagrams of the preceding sort, we can now interpret the epistemic language:

Definition Truth conditions.

$\mathbf{M}, s \models p$	iff	*V makes p true at s*
$\mathbf{M}, s \models \neg\varphi$	iff	*not* $\mathbf{M}, s \models \varphi$
$\mathbf{M}, s \models \varphi \wedge \psi$	iff	$\mathbf{M}, s \models \varphi$ *and* $\mathbf{M}, s \models \psi$
$\mathbf{M}, s \models K_i\varphi$	iff	*for all t with* $s \rightarrow_i t$: $\mathbf{M}, t \models \varphi$
$\mathbf{M}, s \models C_G\varphi$	iff	*for all t that are reachable from s by some finite sequence of* $\rightarrow_i$ *steps* $(i \in G)$: $\mathbf{M}, t \models \varphi$ [22]

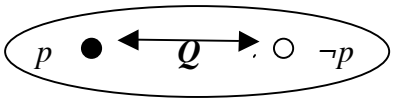
■

Example A model for a question/answer scenario.

Here is how a question answer episode might start (this is just one of many possible initial situations!). In the following diagram, reflexive arrows are presupposed, but not drawn

[22] Thus, we quantify over all sequences like $\rightarrow_1 \rightarrow_2 \rightarrow_3 \rightarrow_1$, etc. For the cognoscenti, in this way, common knowledge acts as a *dynamic logic* modality $[(\cup_{i \in G} \rightarrow_i)^*]\varphi$.

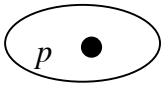
(reflexivity represents the usual assumption that knowledge is truthful). Intuitively, agent ***Q*** does not know whether p, but ***A*** is fully informed about it:



In the black world to the left, the following formulas are true:

$$p, K_A p, \neg K_Q p \wedge \neg K_Q \neg p, K_Q(K_A p \vee K_A \neg p),$$
$$C_{\{Q,A\}}(\neg K_Q p \wedge \neg K_Q \neg p), C_{\{Q,A\}}(K_A p \vee K_A \neg p)$$

This is an excellent situation for ***Q*** to ask ***A*** whether p is the case: he even knows that she knows the answer. Once the answer "Yes" has been given, intuitively, this model changes to the following one-point model where maximal information has been achieved:



Now, of course $C_{\{Q,A\}}p$ holds at the black world. ■

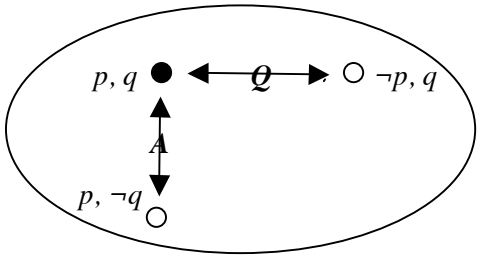
Over our models, an epistemic language sharpens distinctions. For instance, that everyone in a group knows φ is not yet common knowledge, but *universal knowledge* $E_G\varphi$, being the conjunction of all formulas $K_i\varphi$ for all $i \in G$. Now let us broaden things still further.

The above scenarios can also be seen as getting information from any source: e.g., Nature in the case of *observation*. Observation or general learning are congenial to epistemic logic. We just emphasized the conversational setting because it is lively and intuitive.

From semantics to language design There is something conservative to 'giving a semantics': a language is given, and we fit it to some model class. But epistemic models are appealing in their own right as geometrical information structures. And then, there is an issue what language best describes these structures, perhaps changing the given one. To illustrate this design issue, we return to the *social character* of group information.

Example From implicit to explicit group knowledge.

Consider a setting where both agents have information that the other lacks, say as follows:



Here, the black dot on the upper left is the actual world. The most cooperative scenario here is for ***Q*** to tell ***A*** that q is the case (after all, this is something he knows), while ***A*** can tell him that p is the case. Intuitively, this reduces the initial three-point model to the one-point model where $p \wedge q$ is common knowledge. Other three-world examples model other interesting, sometimes surprising conversational settings (cf. van Benthem 2006C). ■

Another way of saying what happens here is that, when ***Q***, ***A*** maximally inform each other, they will cut things down to the *intersection* of their individual accessibility relations. This suggests a new natural notion for groups, beyond the earlier common knowledge:

Definition Distributed knowledge.
Intuitively, a formula φ is *implicit* or *distributed knowledge* in a group, written $D_G\varphi$, when agents could come to see it by pooling their information. More technically, extending our language and the above truth definition, this involves intersection of accessibility relations:

$\boldsymbol{M}, s \models D_G\varphi$ iff for all t with $s \bigcap_{i \in G} \rightarrow_i t$: $\boldsymbol{M}, t \models \varphi$ [23] ■

Intuitively, groups of agents can turn their implicit knowledge into common knowledge (modulo some technicalities) by communication. We will pursue this in Chapters 3, 12.

As we proceed, we will use other extensions of the basic language as needed, including a *universal modality* over all worlds (accessible or not) and *nominals* defining single worlds.

Digression: epistemic models reformulated Though our logic works over arbitrary models, in practice, equivalence relations are common. Our card examples were naturally described as follows: each agent has local states it can be in (say, the card deals it may receive), and worlds are global states consisting of vectors X, Y with each agent in some local state. Then the natural accessibility relation goes via component-wise equality:

$X \rightarrow_i Y$ *iff* $(X)_i = (Y)_i$.

Another case are the models for games in Chapters 10, 14, where worlds are vectors of strategies for players. This is a 'normal form' (Fagin et al. 1995, van Benthem 1996):

Fact Each epistemic model with equivalence relations for its accessibilities
is isomorphic to a sub-model of a vector model.

---

[23] Other definitions of implicit knowledge exist, preserving our later bisimulation invariance. Cf. Roelofsen 2006, van der Hoek, van Linder & Meijer 1999.

Proof Give each agent as its states the equivalence classes of its accessibility relation. This yields an isomorphism since $\rightarrow_i$-equivalence classes of worlds s, t are equal iff $s \rightarrow_i t$. ■

2.3 Validity and axiomatic systems

Validity of formulas φ in epistemic logic is defined as usual in semantic terms, as truth of φ in all models at all worlds. Consequence may be defined through validity of conditionals.

Minimal logic The following completeness result says that the validities over arbitrary models may be described purely syntactically by the following calculus of deduction: [24]

Theorem The valid formulas are precisely the theorems of the *minimal epistemic logic* axiomatized by (a) all valid principles of propositional logic, (b) the definition $<>\varphi \leftrightarrow \neg K \neg \varphi$, (c) modal distribution $K(\varphi \rightarrow \psi) \rightarrow (K\varphi \rightarrow K\psi)$, together with the inference rules of (d) Modus Ponens ('from $\varphi \rightarrow \psi$ and φ, infer ψ') and (e) Necessitation ('if φ is a theorem, then so is $K\varphi$').

Proof The proof for this basic result can be found in any good textbook. Modern versions employ Henkin-style constructions with maximally consistent sets over some finite set of formulas only, producing a finite counter-model for a given non-derivable formula. ■

One axiom of this simple calculus has sparked continuing debate, viz. the distribution law

$$K(\varphi \rightarrow \psi) \rightarrow (K\varphi \rightarrow K\psi)$$

This seems to say that agents' knowledge is closed under logical inferences, and this 'omniscience' seems unrealistic. At stake here is our earlier distinction between semantic and inferential information. Semantically, with processes of observation, an agent who has the hard information that $\varphi \rightarrow \psi$ and that φ also has the hard information that ψ. But in a more finely-grained perspective of syntactic inferential information, geared toward processes of elucidation, Distribution need not hold. We will not join the fray here, but Chapter 5 shows one way of introducing inferential information into basic epistemic logic and Chapter 13 contains a more extensive philosophical discussion.

Internal versus external Another point to note is this. Deductive systems may be used in two modes. They can describe agents' own reasoning inside scenarios, or outside reasoning by theorists about them. In some settings, the difference will not matter: the modeler is one

[24] We will often drop agent subscripts for K-operators when they play no essential role.

of the boys – but sometimes, it may. In this book, we will not distinguish 'first person' and 'third person' perspectives on epistemic logic, as our systems accommodate both. [25]

Stronger epistemic logics and frame correspondence On top of this minimal deductive system, two further steps go hand in hand: helping ourselves to stronger axioms endowing agents with further features, and imposing further structural conditions on accessibility in our models. For instance, here are three more axioms with vivid epistemic interpretations:

$K\varphi \rightarrow \varphi$	*Veridicality*
$K\varphi \rightarrow KK\varphi$	*Positive Introspection*
$\neg K\varphi \rightarrow K\neg K\varphi$	*Negative Introspection*

The former seems uncontroversial (knowledge is in synch with reality), but the latter two have been much discussed, since they assume that, in addition to their logical omniscience, agents now also have capacities of unlimited introspection into their own epistemic states. Formally, these axioms correspond to the following structural conditions on accessibility:

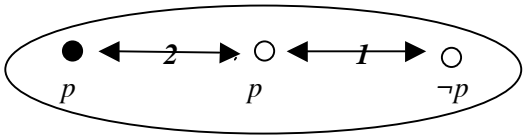
$K\varphi \rightarrow \varphi$	*reflexivity*	$\forall x: x \rightarrow x$
$K\varphi \rightarrow KK\varphi$	*transitivity*	$\forall xyz: (x \rightarrow y \wedge y \rightarrow z) \Rightarrow x \rightarrow z$
$\neg K\varphi \rightarrow K\neg K\varphi$	*euclidity*	$\forall xyz: (x \rightarrow y \wedge x \rightarrow z) \Rightarrow y \rightarrow z$

The term correspondence can be made precise using *frame truth* of modal formulas under all valuations on a model (cf. van Benthem 1984). Powerful results exist matching up axioms with relational conditions: first-order like here, or in higher-order languages. We will occasionally refer to such techniques, but refer to the literature for details.

The complete deductive system with all the above axioms is called *S5*, or *multi-S5* when we have more than one agent. Taking the preceding conditions together, it is the logic of equivalence relations over possible worlds. What may seem surprising is that this logic has no interaction axioms relating different modalities K_i, K_j. But none are plausible:

Example Your knowledge and mine do not commute.

The following model provides a counter-example to the implication $K_1 K_2 p \rightarrow K_2 K_1 p$. Its antecedent is true in the black world to the left, but its consequent is false:



[25] Aucher 2008 is a complete reworking of dynamic epistemic logic in the internal mode.

Such implications only hold when agents have special informational relationships. [26] ■

In between the minimal logic and *S5*, many other logics live, such as *KD45* for belief. In this book, we use examples from the extremes, though our results apply more generally.

Group logic We conclude with a typical additional operator for groups of agents:

Theorem The complete epistemic logic with common knowledge is axiomatized by the following two principles in addition to the minimal epistemic logic, where E_G is the earlier modality for 'everybody in the group knows':

$C_G\varphi \leftrightarrow (\varphi \wedge E_G C_G \varphi)$ *Fixed-Point Axiom*

$(\varphi \wedge C_G(\varphi \rightarrow E_G\varphi)) \rightarrow C_G\varphi$ *Induction Axiom*

These laws are of wider interest. The Fixed-Point Axiom expresses reflexive equilibrium: common knowledge of φ is a proposition p implying φ while every group member knows that p is true. The Induction Axiom says that common knowledge is not just any such proposition, but the largest: technically, a 'greatest fixed-point' (cf. Chapters 3, 4, 10).

This completes our tour of the basics of epistemic logic. Next, we survey some technical themes that will play a role later in this book. Our treatment will be light, and we refer to the literature for details (Blackburn, de Rijke & Venema 2000, van Benthem, to appearB).

2.4 Bisimulation invariance and expressive power

Given the importance of languages in this book, we first elaborate on expressive power and invariance. Expressive strength of a language is often measured by its power of telling models apart by definable properties, or by invariance relations measuring what it cannot distinguish. One basic invariance is *isomorphism*: structure-preserving bijection between models. First-order formulas $\varphi(\boldsymbol{a})$ cannot distinguish between a tuple of objects $\boldsymbol{a}$ in one model $\boldsymbol{M}$, and its image $f(\boldsymbol{a})$ in another model $\boldsymbol{N}$ linked to $\boldsymbol{M}$ by an isomorphism f. [27] This style of analysis also applies to epistemic logic. But first we make a technical connection:

---

[26] In the combined dynamic-epistemic logics of Chapter 3 and subsequent ones, some operator commutation principles will hold, but then between epistemic modalities and action modalities.

[27] Another candidate is *potential isomorphism*: van Benthem 2002B has some extensive discussion. Again we refer to the cited literature for more details in what follows.

First-order translation Viewed as a description of the above epistemic models, our language is weaker than a first-order logic whose variables range over worlds:

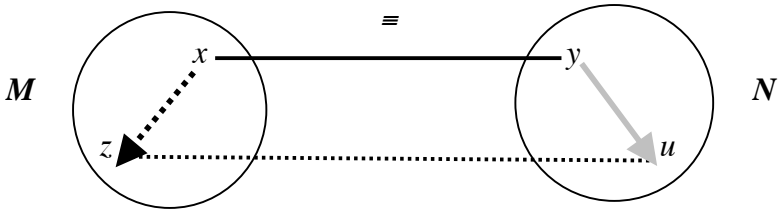
Fact There exists an effective translation from the basic epistemic language into first-order logic yielding equivalent formulas.

Proof This is the well-known Standard Translation. E.g., the epistemic formula $p \wedge K\Diamond q$ goes to an equivalent $Px \wedge \forall y(Rxy \rightarrow \exists z(Ryz \wedge Qz)$, with R a binary predicate symbol for the accessibility relation, and P, Q unary predicates for proposition letters. The first-order formula is a straightforward transcription of the truth conditions for the epistemic one. ■

These modal translations have only special bounded or *guarded* quantifiers over accessible worlds, making them a special subclass of first-order logic (van Benthem 2005B). This shows in powers of distinction. The full first-order language can distinguish a one-point reflexive cycle from an irreflexive *2*-cycle (two non-isomorphic models) - but it should be intuitively clear that these models verify the same epistemic formulas everywhere.

Bisimulation and information equivalence Here is a notion of semantic invariance that fits the modal language like a Dior gown:

Definition A *bisimulation* between two models $\boldsymbol{M}$, $\boldsymbol{N}$ is a binary relation $\leftrightarrows$ between their states s, t such that, whenever $x \leftrightarrows y$, then (a) x, y satisfy the same proposition letters, (b1) if $x\ R\ z$, then there exists a world u with $y\ R\ u$ and $z \leftrightarrows u$, and (b2) the same 'zigzag' or 'back-and-forth clause' holds in the opposite direction. The following diagram shows this:



■

Clause (1) expresses 'local harmony', the zigzag clauses (2) the dynamics of simulation. This is often given a procedural spin: bisimulation identifies processes that run through similar states with similar local choices. Thus, it answers a fundamental question about computation: "When are two processes the same?" [28]

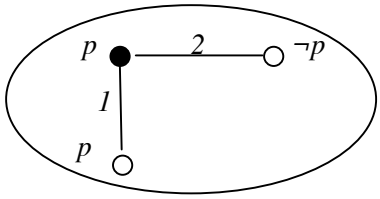
[28] There are different natural notions of identity between processes (cf. van Benthem 1996) and other structures, such as games (cf. van Benthem 2002A and Chapter 10). This diversity in answers is sometimes called Clinton's Principle: *It all depends on what you mean by 'is'*.

Likewise, we can ask:

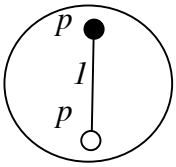
When are two information models the same?

Example Bisimulation-invariant information models.

Bisimulation occurs naturally in epistemic update changing a current model. Suppose that the initial model is like this, with the actual world indicated by the black dot:



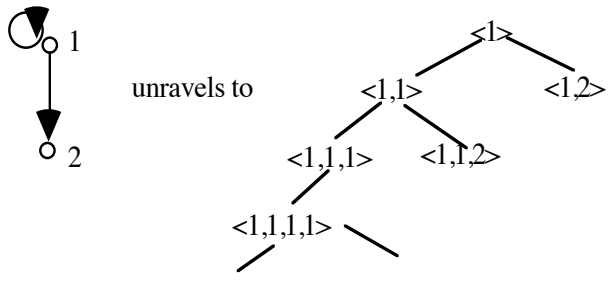
All three worlds satisfy different epistemic formulas. Now, despite her uncertainty, in the actual world, agent *1* does know that *p*, and can say this – updating to the model



But here the two worlds are intuitively redundant, and indeed this information state for the two agents has an obvious bisimulation to just the one-point model

■

Each model has a smallest 'bisimulation contraction' satisfying the same modal formulas. But bisimulation can also make models larger, turning them into *trees* through the method of 'unraveling', providing a nice geometrical normal form (used in Chapter 11):



Thus, informational equivalence can work both ways, depending on what we find useful.

Invariance and definability Some model-theoretic results tie bisimulation closely to truth of modal formulas. Just for convenience, we restrict attention to *finite models* – but this is easily generalized in the theory of modal logic. We formulate results for general relations:

Invariance Lemma The following two assertions are equivalent:

(a) $\boldsymbol{M}, s$ and $\boldsymbol{N}, t$ are connected by a bisimulation,

(b) $\boldsymbol{M}, s$ and $\boldsymbol{N}, t$ satisfy the same epistemic formulas.

Proof The direction from (a) to (b) is by induction on epistemic formulas. In the opposite direction, for finite models, we argue as follows. Let $(\boldsymbol{M}, x)\ Z\ (\boldsymbol{N}, y)$ be the relation which holds if the worlds x and y satisfy the same epistemic formulas. For a start, then, $s\ Z\ t$. Clearly also, Z-related worlds satisfy the same proposition letters. Now, assume that $x\ Z\ y$, and let $x \rightarrow x'$. Suppose for contradiction, that there is no world y' with $y \rightarrow y'$ satisfying the same modal formulas. Then there is a modal formula $\alpha_{y'}$ true in x' and false in y'. Taking the conjunction α of all these formulas $\alpha_{y'}$ for all successors y' of y, we see that α is true in x', and so $\Diamond\alpha$ is true in world x, whence $\Diamond\alpha$ is also true in y, as $x\ Z\ y$. But then there must be a successor y^* of y satisfying α: and this contradicts the construction of α. [29] ■

Semantic versus syntactic information states The epistemic language and bisimulation-invariant structure are two sides of one coin. This is relevant to our view of information states. Looking explicitly, the *epistemic theory* of a world s in a model $\boldsymbol{M}$ is the set of all formulas that are true internally at s about the facts, agents' knowledge of these, and their knowledge of what others know. By contrast, the models $\boldsymbol{M}, s$ themselves locate the same information implicitly in the local valuation of a world plus its pattern of interaction with other worlds. [30] The next result says that these two views of information are equivalent:

State Definition Lemma For each model $\boldsymbol{M}, s$ there exists an epistemic formula β (involving common knowledge) such that the following are equivalent:

(a) $\boldsymbol{N}, t \models \beta$

(b) $\boldsymbol{N}, t$ has a bisimulation $\equiv$ with $\boldsymbol{M}, s$ such that $s \equiv t$

Proof This result is from Barwise & Moss 1996 (our version follows van Benthem 1997). We sketch the proof for equivalence relations $\sim_a$ with existential modalities $\langle a \rangle$, but it works for arbitrary relations. First, any finite multi-*S5* model $\boldsymbol{M}, s$ falls into maximal zones of worlds that satisfy the same epistemic formulas in our language.

---

[29] The lemma even holds for arbitrary epistemic models, provided we take epistemic formulas from a language with arbitrary *infinite* conjunctions and disjunctions.

[30] Compare the way in which category theorists describe a mathematical structure externally by its *connections to other objects* in a category through the available morphisms.

Claim 1 There is a finite set of formulas φ_i $(1 \leq i \leq k)$ defining a partition of the model, and if two worlds satisfy the same φ_i, then they agree on all epistemic formulas.

To see this, take any world s, and take difference formulas $\delta^{s,t}$ between it and any t not satisfying the same epistemic formulas: say, s satisfies $\delta^{s,t}$ and t does not. The conjunction of all $\delta^{s,t}$ is a formula φ_i true only in s and all worlds sharing its epistemic theory. We may assume φ_I also lists all information about proposition letters true and false throughout its zone. We also make a quick observation about uncertainty links between these zones:

\# If any world satisfying φ_I is $\sim_a$-linked to a world satisfying φ_i, then all worlds satisfying φ_I also satisfy $\langle a \rangle \varphi_i$

Next take the following description $\beta_{\mathbf{M}, s}$ of $\mathbf{M}, s$:

(a) all proposition letters and their negations true at s, plus the unique φ_i true at $\mathbf{M}, s$

(b) common knowledge of (b1) the disjunction of all the zone formulas φ_i, (b2) all negations of conjunctions $\varphi_i \wedge \varphi_j$ $(i \neq j)$, (b3) all true implications $\varphi_i \rightarrow \langle a \rangle \varphi_j$ for which situation # occurs, (b4) all true implications $\varphi_i \rightarrow [a] \vee \varphi_j$, with a disjunction $\vee$ over all cases enumerated in (b3).

Claim 2 $\mathbf{M}, s \models \beta_{\mathbf{M}, s}$

Claim 3 If $\mathbf{N}, t \models \beta_{\mathbf{M}, s}$, then there is a bisimulation between $\mathbf{N}, t$ and $\mathbf{M}, s$.

To prove Claim 3, let $\mathbf{N}, t$ be any model for $\beta_{\mathbf{M}, s}$. The φ_i partition $\mathbf{N}$ into disjoint zones Z_i of worlds satisfying these formulas. Now relate all worlds in such a zone to all worlds that satisfy φ_i in the model $\mathbf{M}$. In particular, t gets connected to s. We check that this gives a bisimulation. The atomic clause is clear by construction. But also, the zigzag clauses follow from the given description. (a) Any $\sim_a$-successor step in $\mathbf{M}$ has been encoded in a formula $\varphi_i \rightarrow \langle a \rangle \varphi_j$ that holds everywhere in $\mathbf{N}$, producing the required successor there. (b) Conversely, if there is no $\sim_a$-successor in $\mathbf{M}$, this shows up in the limitative formula $\varphi_i \rightarrow [a] \vee \varphi_j$, which also holds in $\mathbf{N}$, so that there is no excess successor there either. ■

The Invariance Lemma says that bisimulation has a good fit with the modal language. The State Definition Lemma strengthens this to say that each semantic state is captured by one formula. Again this extends to arbitrary models with an infinitary epistemic language.

Example Defining a model up to bisimulation.

Consider the two-world model for our earlier basic question-answer episode. Here is an epistemic formula that defines its φ-state up to bisimulation:

$$\varphi \wedge C_{\{Q, A\}}((K_A\varphi \vee K_A\neg\varphi) \wedge \neg K_Q\varphi \wedge \neg K_Q\neg\varphi) \quad \blacksquare$$

Thus we can switch between syntactic and semantic information states. The latter view will dominate this book, but the reader may want to keep this duality in mind.

2.5 Computation and the complexity profile of a logic

While derivability and definability are the main pillars of logic, issues of *task complexity* form a natural complement. Given that information has to be recognized or extracted to be of use to us, it is natural to ask how complex such extraction processes really are.

Decidability In traditional modal logic, the interest in complexity has gone no further than just asking the following question. Validity in first-order logic is *undecidable*, validity in propositional logic is *decidable*: what about modal logic, which sits in between?

Theorem Validity in the minimal modal logic is decidable. So is that for multi-*S5*.

There are many proofs, exploiting special features of modal languages, especially, their bounded local quantifiers. One method uses the *effective finite model property*: each satisfiable modal formula φ has a finite model whose size can be computed effectively from the length of φ.[31] Validity is decidable for many further modal and epistemic logics.

But things can change rapidly when we consider logics with combinations of modalities, namely, when these show what looks like natural commutation properties:

Theorem The minimal modal logic of two modalities *[1], [2]* satisfying the axiom $[1][2]\varphi \rightarrow [2][1]\varphi$ plus a universal modality *U* over all worlds is undecidable.

The technical reason is that such logics encode undecidable *tiling problems* on the structure *IN x IN* (Harel 1985, Marx 2006). By frame correspondence, the commutation axiom defines a *grid structure* satisfying the following first-order convergence property:

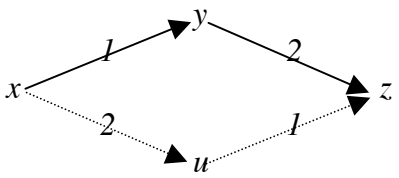
$$\forall xyz: (xR_1y \wedge yR_2z) \rightarrow \exists u: (xR_2u \wedge uR_1z).$$ [32]

Here is a diagram picturing this, creating a cell of the grid:

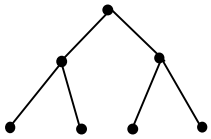
---

[31] This property typically fails for the full language of first-order logic.

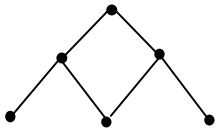
[32] The universal modality has to do with details of the embedding: cf. van Benthem to appearB.



This complexity danger is general, and the following pictures will help you remember it. Modal logics of trees are harmless, modal logics of grids are dangerous!



tree



grid

Modal logics with such axioms occur with agency (Chapter 10, 11), and undecidability is around the corner. But in this chapter, we first look at the fine-structure of decidable tasks.

Computational complexity Complexity theory studies computation time or memory space needed for tasks as a function of input size. Inside the decidable problems, it distinguishes feasible polynomial rates of growth (linear, quadratic, …: the time complexity class ***P***) from non-feasible non-deterministic polynomial time ***NP***, polynomial space (***Pspace***), exponential time (***Exptime***), and ever higher up. For details, see Papadimitriou 1994.

Complexity profile of a logic To really understand how a logical system works, it helps to check the complexity for some basic tasks that it would be used for. Determining validity, or equivalently, *testing for satisfiability* of given formulas, is one of these:

Given a formula φ, determine whether φ has a model.

But there are other, equally important tasks, such as *model checking*:

Given a formula φ and a finite model $(\mathbf{M}, s)$, check whether $\mathbf{M}, s \models \varphi$.

And here is a third key task, that one might call *testing for model equivalence*:

Given two finite models $(\mathbf{M}, s)$, $(\mathbf{N}, t)$, check if they satisfy the same formulas.

Here is a table of the complexity profiles for two well-known classical logics. Entries mean that the problems are in the class indicated, and no lower or higher:

	Model Checking	Satisfiability	Model Comparison
Propositional logic	*linear time (**P**)*	***NP***	*linear time (**P**)*
First-order logic	***Pspace***	*undecidable*	***NP***

Where does the basic modal language fit? Its model checking is efficient. While first-order evaluation has exponential growth via quantifier nesting, there are fast modal algorithms. Next, close-reading decidability arguments helps us locate satisfiability. Finally, testing for modal equivalence, or for the existence of a bisimulation, has turned out efficient, too:

Fact The complexity profile for the minimal modal logic is as follows:

Model Checking	Satisfiability	Model Comparison
P	***Pspace***	***P***

For epistemic logic, results are similar. But *S5* has a difference. Single-agent satisfiability is in ***NP***, as *S5* allows for a normal form without iterated modalities. But with two agents, satisfiability jumps back to ***Pspace***: social life is more complicated than being alone.

Complexity results are affected by the expressive power of a language. When we add our common knowledge modality $C_G\varphi$ to epistemic logic, the above profile changes as follows:

Model Checking	Satisfiability	Model Comparison
P	***Exptime***	***P***

The Balance: expressive power versus computational effort Logic has a Golden Rule: what you gain in one desirable dimension, you lose in another. Expressive strength means high complexity. Thus, it is all about striking a balance. First-order logic is weaker than second-order logic in defining mathematical notions. But its poverty has a reward, viz. the axiomatizability of valid consequence, and useful model-theoretic properties such as the Compactness Theorem. Many modal logics are good compromises lower down this road. They become decidable, or even if not: they have more perspicuous proofs systems.

2.6 Games for logical tasks

Computation is not just a routine chore: procedures are a fundamental theme in their own right. This comes out well with *game versions* of logical tasks (van Benthem 2007D):

Evaluation games We cast the process of evaluating modal formula φ in model $(\mathbf{M}, s)$ as a *two-person game* between a Verifier, claiming that φ is true, and a Falsifier claiming that φ is false. The game starts at some world *s*. Each move is dictated by the main operator of the formula at hand (the total length is again bounded by its modal depth):

Definition The modal evaluation game $game(\mathbf{M}, s, \varphi)$.

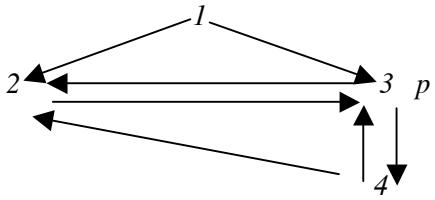
The rules of the games endow the logical operators with an interactive dynamic meaning:

atom p	test p at s in $\boldsymbol{M}$: if true, then V wins – otherwise, F wins
disjunction	V *chooses* a disjunct, and play continues with that
conjunction	F chooses a conjunct, and play continues with that
$<>\varphi$	V picks an R-successor t of the current world; play continues with φ at t
$[]\varphi$	F picks an R-successor t of the current world; play continues with φ at t.

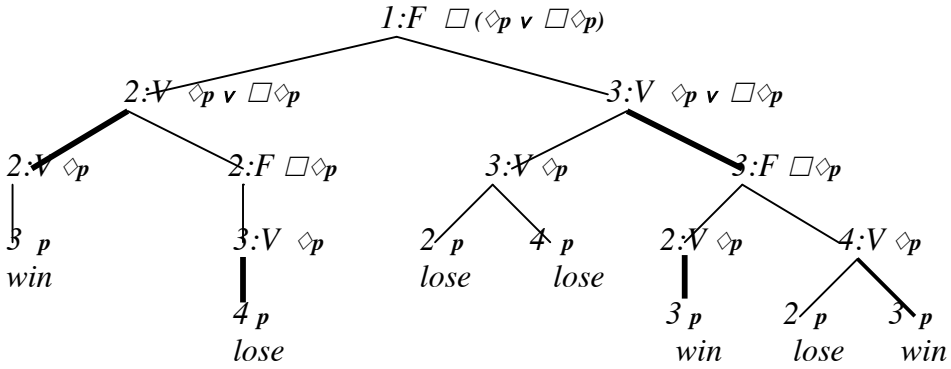
A player also loses when (s)he must pick a successor, but cannot do so. ■

Example A complete game tree.

We give an illustration in general modal logic. Here is the complete game tree for the modal formula $\Box(\Diamond p \vee \Box\Diamond p)$ played starting at state *1* in the following model:



We draw game nodes, plus the player which is to move, plus the relevant formula. Bottom leaves have 'win' if Verifier wins (the atom is true there), 'lose' otherwise:



Each player has three winning runs, but the advantage is for Verifier, who can always play to win, whatever Falsifier does. Her *winning strategy* is marked by the bolder lines. ■

A strategy can encode subtle behaviour: Verifier must hand the initiative to Falsifier at state *3* on the right if she is to win. Some background is found in Chapter 10. The reason why Verifier has a winning strategy is that she is defending a true statement:

Fact For any modal evaluation game, the following two assertions are equivalent:

(a) formula φ is true in model $\boldsymbol{M}$ at world s,

(b) player V has a winning strategy in *game*($\boldsymbol{M}$, s, φ).

Proof The proof is a straightforward induction on the formula φ. This is a useful exercise to get a feel for the workings of strategies, and the game dynamics of the modalities. ■

Model comparison games The next game provides fine-structure for bisimulation between models *(**M**, s)*, *(**N**, t)*. They involve a Duplicator (claiming that there is an analogy) and Spoiler (claiming a difference), playing over pairs *(x, y)* in the models.

Definition Modal comparison games.

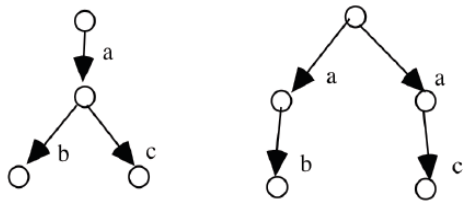
In each round of the game, Spoiler chooses a model, and a world *u* that is a successor of *x* or *y*, and Duplicator responds with a corresponding successor *v* in the other model. Spoiler wins if *u, v* differ in their atomic properties, or Duplicator cannot find a successor in 'her' model. The game continues over some finite number of rounds, or infinitely. ■

There is a tight connection between these games and modal properties of the two models:

Fact (a) Spoiler's winning strategies in a *k*-round game between ***M**, s* and ***N**, t* match exactly the modal formulas of *operator depth k* on which *s, t* disagree.
(b) Duplicator's winning strategies in the *infinite* round game between ***M**, s* and ***N**, t* match the bisimulations between ***M***, ***N*** linking *s* to *t*.

It is not hard to prove this result generally, but we merely give some illustrations.

Example 'Choosing now or later'.
Consider the game between the following two models starting from their roots:



Spoiler can win the comparison game in *2* rounds, with several strategies. One stays inside the same model, exploiting the modal difference $\langle a\rangle(\langle b\rangle T \wedge \langle c\rangle T)$ of depth *2*. [33] Another winning strategy for Spoiler switches models, using the modal formula $[a]\langle b\rangle T$ that contains only *2* operators in all. The switch is signaled by the change in modalities. ■

This concludes our sketch of games that show logic *itself* in a dynamic interactive light. [34] We discuss this issue of logics as dynamic procedures at greater length in (van Benthem, to

[33] Here *'T'* stands for the *always true* formula.

[34] Other nice logic games do tasks like model construction or proof (van Benthem 2007D).

appear). For here, note also that the above game trees are themselves modal models. This suggests applications of modal logic to game theory, as we will see in Chapter 10.

2.7 Conclusion

We have presented epistemic logic in a new manner. First, we shook off some dust, and emphasized the themes of hard semantic information and modeling multi-agent scenarios. In doing so, we dropped the claim that the *K*-operator stands for knowledge in any philosophical sense, making it a statement about current semantic information instead. This is a radical shift, and the issue of what real knowledge is now shifts to the analysis of further aspects of that notion such as explicit awareness (Chapter 5), belief (Chapter 7) and dynamic stability under new information (cf. the epistemological passages in Chapter 13).

Next we presented the usual basics of the formal system, plus some newer special topics. Together, these illustrated three key themes: (a) expressive power of definition, (b) inferential power of deduction, and (c) computational task performance, leading to a suggestive interactive embodiment of the logic itself in dynamic games.

2.8 Further directions and open problems

There are many further issues about knowledge, information, and logic. A few have been mentioned, such as connections with dynamic logics, or the contrast between first and third person perspectives. We mention a few more, with pointers to the literature.

Topological models Relational possible worlds models are a special case of a more general and older semantics for modal logic. Let $\mathbf{M}$ be a topological space $(X, \mathbb{O}, V)$ with points X, a topology of open sets $\mathbb{O}$, and a valuation V. This supports a modal logic:

Definition Topological models.
φ is true at point s in $\mathbf{M}$, written $\mathbf{M}, s \models \Box\varphi$, if s is in the *topological interior* of the set $[[\varphi]]^{\mathbf{M}}$: the points satisfying φ in $\mathbf{M}$. Formally, $\exists O \in \mathbb{O}: s \in O \;\&\; \forall t \in O: \mathbf{M}, t \models \varphi$. ■

Typical topologies are metric spaces like the real numbers, or trees with sets closed in the tree order as opens (essentially, our relational models). Modal axioms state properties of interior: $\Box\varphi \rightarrow \varphi$ is inclusion, $\Box\Box\varphi \leftrightarrow \Box\varphi$ idempotence, and $\Box(\varphi \wedge \psi) \leftrightarrow \Box\varphi \wedge \Box\psi$ closure of opens under intersections. But infinitary distribution, relationally valid, fails:

Example $\Box \bigwedge_{i \in I} p_i \leftrightarrow \bigwedge_{i \in I} \Box p_I$ fails on metric spaces.

Interpret p_i as the open interval $(-1/i, +1/i)$, all $i \in N$. The $\Box p_i$ all denote the same interval,

and the intersection of all these intervals is *{0}*. But the expression $\Box \bigwedge_{i \in I} p_i$ denotes the topological interior of the singleton set *{0},* which is the empty set $\varnothing$. ■

All earlier techniques generalize, including a topological bisimulation related to continuous maps, plus a vivid matching game. The Handbook Aiello, Pratt & van Benthem, eds., 2007 has chapters on the resulting theory. Van Benthem & Sarenac 2005 use these models for epistemic logic, with multi-agent families of topologies closed under operations of group formation. They exploit the failure of infinitary distribution to show the following separation of iterative and fixed-point views of common knowledge (cf. Barwise 1985):

Theorem On topological models, common knowledge defined through countable iteration is strictly weaker than common knowledge defined as a greatest fixed-point for the above equation $C_G \varphi \leftrightarrow \varphi \wedge E_G C_G \varphi$.

In fact, the standard product topology on topological spaces models a form of group knowledge stronger than both, as 'having a shared situation'. Given the epistemic interest in such richer models, there is an open problem of extending the dynamic logics of this book from relational to topological models.

Neighbourhood models A further generalization of dynamic epistemic logic extends the topological setting to abstract *neighbourhood models* where worlds have a family of subsets as their 'neighbourhoods'. Then a box modality $\Box\varphi$ is true at a world *w* if at least one of *w*'s neighbourhoods is contained in *[[φ]]* (cf. van Benthem, to appearB, Hansen, Kupke & Pacuit 2008). The resulting modal base logic has $\Box$ upward monotone, but not distributive over $\wedge$ or $\vee$. The epistemic update rules of our later Chapters 3, 4 extend to neighbourhood models, as long as the set of relevant epistemic events is finite (a generalization of the standard product topology works; cf. the seminar paper Leal 2006). Zvesper 2010 has further developments, motivated by an analysis of some key results in game theory (see Chapter 10). But an elegant general approach remains a challenge.

Tandem View: translation and correspondence This book will mostly use modal logics. But there is the option of translating these to first-order of higher-order logics. Then, our whole theory becomes embedded in *fragments of classical logical systems*. Which ones?

Internal versus external logical views of knowledge Epistemic logic builds on classical propositional logic, adding explicit knowledge operators. By contrast, *intuitionistic logic*

treats knowledge by 'epistemic loading' of the interpretation of standard logical constants like negation and implication. Van Benthem 1993, 2009E discuss the contrast between the two systems, and draw comparisons. As will be shown in Chapter 13, the dynamic content of intuitionistic logic involves observational update (Chapter 3), awareness-raising actions (Chapter 5), as well as procedural information in the sense of Chapter 11. There is a general question of an intuitionistic version for the dynamics explored in this book.

From propositional to predicate logic We have surveyed propositional epistemic logic. But many issues of knowledge have to do with information about objects, such as knowing the location of the treasure, or knowing a method for breaking into a house. This book will not develop *quantificational counterparts*, but it is an obvious next stage throughout.

Chapter 3 DYNAMIC LOGIC OF PUBLIC OBSERVATION

Having laid the groundwork of semantic information and agents' knowledge in Chapter 2, we will now study dynamic scenarios where information flows and knowledge changes by acts of observation or communication. This chapter develops the simplest logic in this realm, that of public announcement or observation of hard information that we can trust absolutely. This simple pilot system raises a surprising number of issues, and its design will be a paradigm for all that follows in the book. We start with motivating examples, then define the basic logic, explore its basic properties, and state the general methodology coming out of this. We then sum up where we stand. At this stage, the reader could skip to Chapter 4 where richer systems start. What follows are further technical themes, open problems (mathematical, conceptual, and descriptive), and a brief view of key sources.

3.1 Intuitive scenarios and information videos

Recall the 'Restaurant scenario of Chapter 1. The waiter with the three plates had a range of 6 possibilities for a start, this got reduced to 2 by the answer to his first question, and then to 1 by the answer to the second. Information flow of this kind means stepwise range reduction. This ubiquitous view can even be seen in propositional logic:

Throwing a party The following device has been used in my courses since the early 1970s. You want to throw a party, respecting incompatibilities. You know that (a) John comes if Mary or Ann does, (b) Ann comes if Mary does not come, (c) If Ann comes, John does not. Can you invite people under these constraints? Logical inference might work as follows:

> By (c), if Ann comes, John does not. But by (a), if Ann comes, John does. This is a contradiction, so Ann does not come. But then, by (b), Mary comes. So, by (a) once more, John must come. Indeed a party {John, Mary} satisfies all three requirements.

This shows the usual propositional rules at work, and the power of inference to get to a goal. But here is a dynamic take on the informational role of the premises that takes things more slowly. At the start, no information was present, and all 8 options remained:

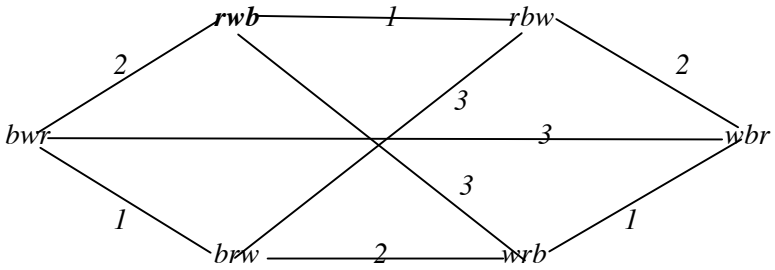
(1) *{MAJ, MA-J, M-AJ, M-A-J, -MAJ, -MA-J, -M-AJ, -M-A-J}* *8 items*

Now the three given premises *update* this initial information state, by removing options incompatible with them. In successive steps, (a), (b), (c) give the following reductions:

(a) ***(M or A) → J***	new state	*{MAJ, M-AJ, -MAJ, -M-AJ, -M-A-J}*	*5 items*
(b) ***not-M → A***	new state	*{MAJ, M-AJ, -MAJ}*	*3 items*
(c) ***A → not-J***	new state	*{M-AJ}*	*1 item*

This resembles the information flow in games like Master Mind (van Benthem 1996), where information about some arrangement of coloured pegs comes in round by round.

Card games The same simple mechanism works in multi-agent settings like *card games*, where it performs sophisticated updates involving what agents know about each other. Recall the 'Three Cards' from Chapter 2, where the following scenario was played during the *NEMO* Science lecture from the Introduction. Cards *red, white, blue* are dealt to players: *1, 2, 3,* one for each. Each player sees his own card only. The real distribution over *1, 2, 3* is red, white, blue (***rwb***). This was the epistemic information model



Now the following two conversational moves take place:

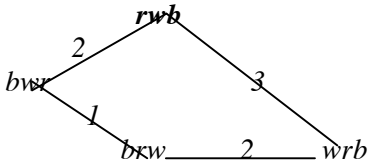
2 asks 1	"Do you have the blue card?"
1 answers truthfully	"No".

Who knows what then? Here is the effect in words:

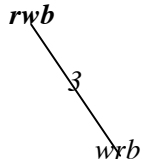
> Assuming the question is sincere, *2* indicates that she does not know the answer, and so she cannot have the blue card. This tells *1* at once what the deal was. But *3* does not learn, since he already knew that *2* does not have blue. When *1* says she does not have blue, this now tells *2* the deal. *3* still does not know even then.

We now give the updates in the diagram, making all these considerations geometrically transparent. Here is a concrete 'update video' of the successive information states:

After *2*'s question:



After *1*'s answer:



We see at once in the final diagram that players *1, 2* know the initial deal now, as they have no uncertainty lines left. But *3* still does not know, given her remaining line, but she does know that *1, 2* know – and in fact, the latter is common knowledge.

Similar analyses exist by now for other conversation scenarios, and for a wide variety of puzzles and games. In particular, one can also model the scenario where *2*'s question is not informative, with an alternative update video whose description we leave to the reader.

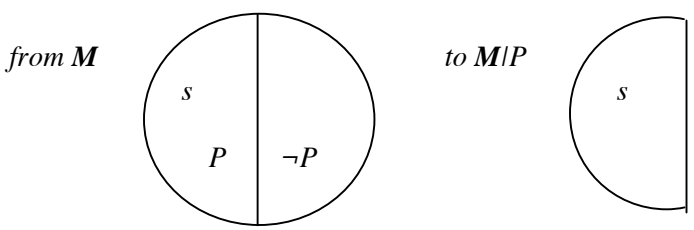
3.2 Modeling informative actions: update by hard information

Our task is now to 'dynamify' the notions of Chapter 2 to deal with these examples. We already have the statics in place, in the form of our earlier epistemic models, described by the epistemic language. Now we need an explicit account of the actions that transform these states, and an extension of the epistemic language to define these explicitly, and reason about them. But what are these basic actions of information flow? In what follows, we will mostly call them *public announcements*, as this term has become wide-spread. But this conversational phrasing reflects just one way of thinking about these basic acts, and equally good cases are *public observation* (the same for every agent), or learning in a more general sense. Indeed, even 'action' may be too specialized a term, and my preference is to think of informational *events*, whether or not with an actor involved. Now, such events come with different force, stronger or weaker. In this chapter, we study the simplest phenomenon: events of *hard information* producing totally trustworthy facts.

Semantic update for hard information Here is what seems a folklore view of information flow: new information eliminates possibilities from a current range. More technically, public announcements or observations *!P* of true propositions *P* yield 'hard information' that changes the current model irrevocably, discarding worlds that fail to satisfy *P*:

Definition Updating via definable submodels.
For any epistemic model $\boldsymbol{M}$, world s, and formula P true at s, the model $(\boldsymbol{M}|P, s)$ ($\boldsymbol{M}$ *relativized to P at s*) is the sub-model of $\boldsymbol{M}$ whose domain is the set $\{t \in \boldsymbol{M} \mid \boldsymbol{M}, t \models P\}$. ■

Drawn in a simple picture, such an update step *!P* goes



These diagrams of a jump from one model to another [35] are useful in visualizing arguments about validity of logical principles in this setting. These principles are not entirely obvious. Crucially, truth values of formulas may change in the update depicted here: [36] most notably, since agents who did not know that *P* now do after the announcement. This makes reasoning about information flow more subtle than just a simple conditionalization. The best way of getting clear on such issues is, of course, the introduction of a logical system.

The Muddy Children This simple update mechanism explains the workings of many knowledge puzzles, one of which has become an evergreen in the area, as it packs many key topics into one simple story. It occurred in Chapter 1, but we repeat it here. A complete discussion of all its features must wait until our subsequent chapters.

Example Muddy Children (Fagin et al. 1995, Geanakoplos 1992).
After playing outside, two of three children have mud on their foreheads. They can only see the others, so they do not know their own status. (This is an inverse of our card games.) Now their Father says: "At least one of you is dirty". He then asks: "Does anyone know if he is dirty?" Children answer truthfully. As questions and answers repeat, what happens?

> Nobody knows in the first round. But in the next round, each muddy child can reason like this: "If I were clean, the one dirty child I see would have seen only clean children, and so she would have known that she was dirty at once. But she did not. So I must be dirty, too!" ■

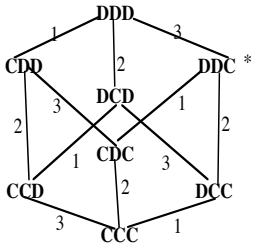
In the initial model, eight possible worlds assign ***D*** or ***C*** to each child. A child knows about the others' faces, not her own, as reflected in the accessibility lines in the diagrams below. Now, the successive assertions made in the scenario update this information:

Example, continued Updates for the muddy children.
Updates start with the Father's public announcement that at least one child is dirty. This simple communicative action merely eliminates those worlds from the initial model where the stated proposition is false. I.e., *CCC* disappears:

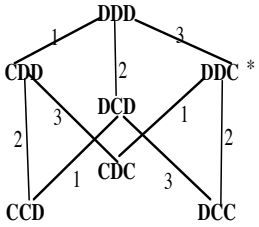
---

[35] Note that an update action *!P* as described here may be seen as a *partial function* on epistemic models: it is only executable when $\boldsymbol{M}, s \models P$, and in that case, it produces a unique new value.

[36] That is why we did not write *'P'* under the remaining zone in the model to the right.

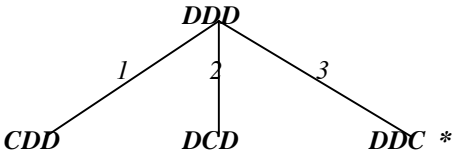


from



to

When no one knows his status, the bottom worlds disappear:

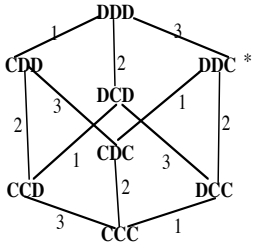


The final update is to ***DDC **** ■

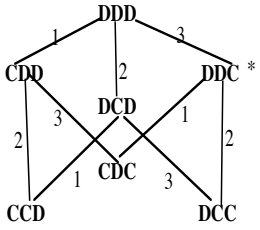
In this model sequence, domain size decreases stepwise: *8, 7, 4, 1*. More generally, with *k* muddy children, *k* rounds of stating the same simultaneous ignorance assertion "I do not know my status" by everyone yield common knowledge about which children are dirty. A few more simultaneous assertions by those who now know achieve common knowledge of the complete actual distribution of the ***D*** and ***C*** for the whole group.

The same setting also analyzes the effects of changes in the procedure. For instance, a typical feature of dynamic actions is that their order of execution matters to the effects:

Example, still continued Dynamic effects of speaking order.
Moving from simultaneous to sequential announcement, the update sequence is quite different if the children speak in turn. The first update is as before, ***CCC*** disappears:



from



to

When the first child says it does not know, *DCC* is eliminated. Then the second child knows its status. Saying this takes out all worlds but *DDC*, *CDC*. In the final model, it is common knowledge that *2, 3* know, but *1* never finds out through epistemic assertions. ■

This is much more than an amusing puzzle. It raises deep issues of information flow that will return in this book. Just consider the following features just below the surface:

Enabling actions. The procedure is jump-started by the Father's initial announcement. No update happens if the further procedure is run on the initial cube-like *8*-world model. Thus, internal communication only reaches the goal of common knowledge after some external information has *broken the symmetry* of the initial diagram. How general is this?

Self-refuting assertions. Children truly state their ignorance, but the last announcement of this fact leads to knowledge of their status, reversing its own truth. How can that be?

Iteration and update evolution. The Father gives an instruction that gets repeated to some limit. We will see that this repetition is a very common scenario, also in games (cf. Chapter 15), and thus, it is of interest to explore universes of evolution as updates get repeated.

Program structures. The scenario involves many program constructions: children follow a rule "If you know your status, say so, else, say you do not" (*IF THEN ELSE*), there is sequential composition of rounds (*;*), we already noted iterations (*WHILE DO*), and the fact that the children speak simultaneously even adds a notion of *parallel composition*.

This combination of examples and deeper issues shows that public announcement of hard information is an update mechanism with hidden depths, and we now proceed to describe it more explicitly in a logical system that can speak explicitly about the relevant events.

3.3 Dynamic logic of public announcement: language, semantics, axioms

First we must bring the dynamics of these successive update steps into a suitable extension of our static epistemic logic. Here is how: [37]

Definition Language and semantics of public announcement.
The language of *public announcement logic PAL* is the epistemic language with added action expressions, as well as dynamic modalities for these, defined by the syntax rules:

Formulas	*P*:	$p \mid \neg\varphi \mid \varphi\wedge\psi \mid K_i\varphi \mid C_G\varphi \mid [A]\,\varphi$ [38]
Action expressions	*A*:	$!P$

The language is interpreted as before in Chapter 2, while the semantic clause for the new dynamic action modality is forward-looking among models as follows:

$\boldsymbol{M}, s \models [!P]\varphi$ iff *if* $\boldsymbol{M}, s \models P$, *then* $\boldsymbol{M}|P, s \models \varphi$ ■

[37] Van Benthem 2006C, D are more extensive surveys of *PAL* and its technical properties.

[38] A convenient existential action modality $\langle A\rangle\varphi$ is defined as usual as $\neg[A]\neg\,\varphi$.

This language allows us to make typical assertions about knowledge change such as

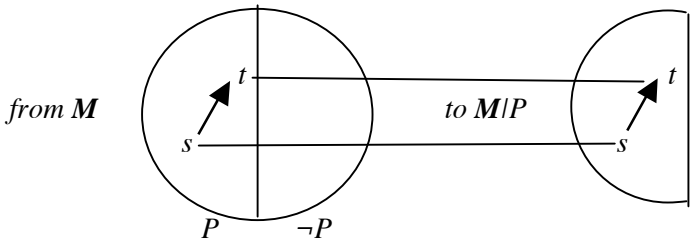
$[!P]K_i\varphi$

that states what an agent i will know after having received the hard information that P. This one formula of dynamified epistemic logic neatly high-lights the combination of ideas from diverse fields that come together here. The study of speech acts $!P$ was initiated in linguistics and philosophy, that of knowledge assertions $K_i\varphi$ in philosophical logic and economics. And the dynamic effect modality $[]$ combining these actions and assertions into a new formal language comes from program logics in computer science. [39]

Axioms and completeness Reasoning about information flow in public update revolves around the formula $[!P]K_i\varphi$. In particular, we need to analyze the dynamic *recursion equation* driving the informational process, that relates the new knowledge to the old knowledge the agent had before the update took place. Here is the relevant principle:

Fact The following equivalence is valid for *PAL*: $[!P]K_i\varphi \leftrightarrow (P \rightarrow K_i(P \rightarrow [!P]\varphi))$.

Proof This can be verified using the above truth clauses, with the above diagrams for concreteness. Compare the models $(\boldsymbol{M}, s)$ and $(\boldsymbol{M}|P, s)$ before and after the update:



[40]

The formula $[!P]K_i\varphi$ says that, in $\boldsymbol{M}|P$, all worlds t $\sim_i$-accessible from s satisfy φ. The corresponding worlds t in $\boldsymbol{M}$ are those $\sim_i$-accessible from s *that satisfy* P. As truth values of formulas may change in an update step, the right description of these worlds in $\boldsymbol{M}$ is not that they satisfy φ (which they do in $\boldsymbol{M}|P$), but rather $[!P]\varphi$: they *become* φ after the update. Finally, $!P$ is a partial function: P must be true for its public announcement to be

[39] *A point of notation*. I write $[!P]\varphi$ to stress the different roles of the announced proposition P and the postcondition φ. Since P and φ are from the same language, the more popular notation $[!\psi]\varphi$ will be used as well. Some literature even has a variant $[\psi]\varphi$ suppressing the action marker $!$.

[40] Note that an update action $!P$ as described here may be seen as a *partial function* on epistemic models: it is only executable when $\boldsymbol{M}, s \models P$, and in that case, it produces a unique new value.

executable. Thus, we make our assertion on the right provided that *!P* is executable, i.e., *P* is true. Putting this together, $[!P]K_i\varphi$ says the same as $P \rightarrow K_i(P \rightarrow [!P]\ \varphi)$. [41] ■

We will discuss this key principle of information flow in more detail later. Here is how it functions in a complete calculus of public announcement. We state a result with a degree of freedom, since we do not care about the precise underlying epistemic base logic:

Theorem *PAL* without common knowledge is axiomatized completely by the laws of epistemic logic over our static model class plus the following *recursion axioms*: [42]

$[!P]q \quad \leftrightarrow \quad P \rightarrow q$ *for atomic facts* q

$[!P]\neg\varphi \quad \leftrightarrow \quad P \rightarrow \neg[!P]\varphi$

$[!P](\varphi \wedge \psi) \quad \leftrightarrow \quad [!P]\varphi \wedge [!P]\psi$

$[!P]K_i\varphi \quad \leftrightarrow \quad P \rightarrow K_i(P \rightarrow [!P]\varphi)$ [43]

Proof First, consider soundness. The first axiom says that update actions do not change the ground facts about worlds. The negation axiom interchanging $[]\neg$ and $\neg[]$ is a special law of modal logic expressing that update is a partial function. The conjunction axiom is always valid. And we have discussed the crucial knowledge axiom already.

Next, we turn to completeness. Suppose that some formula φ of *PAL* is valid. Start with some innermost occurrence of a dynamic modality in a sub-formula $[!P]\psi$ in φ. Now the axioms allow us to push this modality $[!P]$ through Boolean and epistemic operators in ψ until it attaches only to atoms, where it disappears completely thanks to the base axiom. Thus, we get a provably equivalent formula where $[!P]\psi$ has been replaced by a purely epistemic formula. Repeating this process until all dynamic modalities have disappeared, there is a formula φ' provably equivalent to φ. Since φ', too, is valid, it is provable in the base logic, which is complete by assumption, and hence, so is the formula φ itself. ■

Example Announcing an atomic fact makes it known.

Indeed, $[!q]Kq \leftrightarrow (q \rightarrow K(q \rightarrow [!q]q)) \leftrightarrow (q \rightarrow K(q \rightarrow (q \rightarrow q))) \leftrightarrow (q \rightarrow KT) \leftrightarrow T$. [44] ■

---

[41] The consequent simplifies to the equivalent formula $P \rightarrow K_i[!P]\varphi$ usually found in the literature.

[42] Recursion axioms are usually called 'reduction axioms'. But the latter term sounds as if reducing dynamics away is the main point of *PAL* – whereas reduction is a two-edged sword: see below.

[43] Some readers may find equivalent formulations with existential dynamic modalities $\langle !P\rangle$ easier to read. In this book, we will switch occasionally to the latter for reasons of convenience.

Example Diamonds and boxes are close.

For a partial function, a modal diamond and a modal box state almost the same. Here is how this shows in *PAL*: $\langle !P\rangle\varphi \leftrightarrow \neg[!P]\neg\varphi \leftrightarrow \neg(P \rightarrow \neg[!P]\varphi) \leftrightarrow P \wedge [!P]\varphi$. ■

This concludes the introduction of the first dynamic logic of this book. *PAL* is a natural extension of epistemic logic for semantic information that agents have, but it can also talk about events that change this information, and the resulting changes in knowledge. So we have made good on one promise in our program. We now explore things a bit further.

3.4 A first exploration of planet *PAL*

The simple *PAL* system is remarkable in several ways, both technical and conceptual – and it is surprising how many issues of general interest attach to the above calculus.

Descriptive scope *PAL* describes what single agents learn in puzzles like Master Mind, but also multi-agent settings like the Three Cards. It also works in sophisticated scenarios like the Muddy Children that will return in Chapter 11, and the analysis of solution procedures in game theory (Chapter 15). Further examples are speech act theories in philosophy, linguistics, and agent theory in computer science. [45] Of course, *PAL* also has its limits. To mention just one, puzzles or conversations may refer explicitly to the *epistemic past*, as in saying "What you said just now, I knew already". This calls for a past-looking version that does not eliminate worlds for good, but keeps a record of the update history until now (see below, and Chapter 11). Moreover, many realistic scenarios involve partial observation and different information flow for different agents, which will be the topic of our next chapter. Our interest in *PAL* is because it is the first word, not the last.

A lense for new phenomena: Moore sentences, learning and self-refutation *PAL* also helps us see new phenomena. Public announcement of atomic facts p makes them common knowledge. This is often thought to be just the point of speech acts of public assertion. But what guarantees that this is so? Indeed, one must be careful. It is easy to see that the following principle is valid for purely *factual formulas* φ without epistemic operators:

$$[!\varphi]C_G\varphi$$

---

[44] Like earlier, '*T*' stands for the *always true* proposition.

[45] Speech act theory (Searle & Vanderveken 1985) has specifications for successful assertions, questions, or commands. These insights are valuable to applying *PAL* in real settings.

But announcements $!\varphi$ of epistemic truths need not result in common knowledge of φ. A simple counter-example are *Moore-type sentences* that can be true, but never known. For instance, in one of our question-answer scenarios in Chapter 2, let the answerer ***A*** say truly

$p \wedge \neg K_Q p$ "p, but you don't know it"

This very utterance removes the questioner ***Q***'s lack of knowledge about the fact p, and thus makes its own content false. Hence, announcing Moore sentences leads to knowledge of their negation. This switch may seem outlandish, but with the Muddy Children, repeated assertions of ignorance eventually led to knowledge in a last round. Similar beneficial reversals are known from game theory (Dégrémont & Roy 2009).

These switching cases have philosophical relevance. In Chapter 13, we will discuss the Fitch Paradox in Verificationism, the view that every true assertion can become known to us. This thesis must be qualified, as Moore examples show, and following van Benthem 2004B, we will show how this dynamic epistemic logic meets epistemology here. [46] [47]

Making time explicit These examples also highlight a temporal peculiarity of our system. Saying that *PAL* is about learning that φ is ambiguous between: (a) φ *was* the case, before the announcement, and (b) φ *is* the case after the announcement. For worlds surviving in the smaller updated model, factual properties do not change, but epistemic properties may. Making this temporal aspect implicit, and helping ourselves to an ad-hoc *past operator* $Y\varphi$ for 'φ was true at the preceding stage' (Yap 2006, Sack 2008), here is a principle about knowledge, and indeed common knowledge, that does hold in general for all assertions:

$[!\varphi]C_G Y\varphi$

It always becomes common knowledge that φ *was* true at the time of the announcement. [48]

---

[46] For a taxonomy of 'self-fulfilling' and 'self-refuting' statements, see Chapter 15.

[47] A broader issue here is *learnability*: how can we come to know assertions by any means? Define a new modality $\boldsymbol{M}, s \models \langle learn \rangle \varphi$ iff there is an formula P with $\boldsymbol{M}, s \models \langle !P \rangle \varphi$. Prefixing a negation, this can also express that we *cannot* come to know a proposition. Van Benthem 2004B asked if *PAL* plus the modal operator $\langle learn \rangle$ is decidable. Balbiani et al. 2007 proved that it is axiomatizable, but French & van Ditmarsch 2008 then showed that it is undecidable.

[48] Factual formulas φ satisfy an implication $Y\varphi \rightarrow \varphi$ from 'then' to 'now'. Epistemic ones need not.

Iterated assertions or observations In what sense is *PAL* also a logic of conversation or experimental procedure? We will discuss this issue in Chapters 10 and 11, but here is a first pointer. It may have seemed to the reader that the *PAL* axioms omitted one recursive case. Why is there no axiom for a combination of assertions

$$[!P][!Q]\varphi?$$

The reason is that our reduction algorithm started from innermost occurrences of dynamic modalities, always avoiding this case. Still, we have the following observation that tells us how to achieve the effect of saying two consecutive things by saying just one:

Fact The formula $[!P][!Q]\varphi \leftrightarrow [!(P \wedge [!P]Q)]\varphi$ is valid. [49]

The proof is immediate if you think about what this says. The embedding inside the box uses the mutual recursion between actions and formulas in the inductive *PAL* syntax.

The agents behind PAL: observation and memory Static epistemic logic highlights assumptions that have sparked debate. As we saw in Chapter 2, distribution $K(\varphi \rightarrow \psi) \rightarrow (K\varphi \rightarrow K\psi)$ has been read in terms of agents' powers of *inference*, and the *.4* axiom $K\varphi \rightarrow KK\varphi$ as positive *introspection*. In Chapters 5, 13 we will question both for explicit knowledge, but this chapter sticks with semantic information. Even so, *PAL* does put up a new principle for scrutiny that has been much less-discussed, viz. the recursion axiom

$$[!P]K\varphi \leftrightarrow (P \rightarrow K[!P]\varphi) \qquad \textit{Knowledge Gain}$$

What is remarkable about this principle is how it interchanges knowledge after an event with knowledge before that event. What does this mean in terms of agent powers? This is slightly easier to see when we adopt a more abstract stance. Consider the formula

$$K[a]\varphi \leftrightarrow [a]K\varphi$$

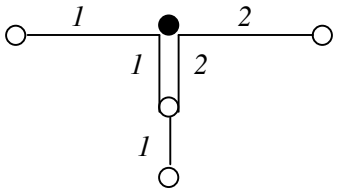
This says that I know now that action a will produce effect φ if and only if, after action a has occurred, I know that φ. For many actions a and assertions φ, this equivalence is fine. But there are counter-examples. For instance, I know now already that, after drinking, I get

[49] Cf. van Benthem 1999B. This is like a colleague who used to monopolize faculty meetings by saying "I say *P*, and then you will say *Q* to that, and then I will say *R* to *Q*", etcetera.

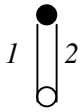
boring. But alas, after drinking, I do not know that I am boring. [50] At stake here are two new features of agents: their powers of *memory*, and of *observation* of relevant events. Drinking impairs these, [51] while public announcements respect both. We will discuss issues of Perfect Recall more fully in Chapter 11, but my point is that powers of observation and memory are just as crucial to epistemology as powers of deduction or introspection.

'Tell All': optimal communication While much of epistemology is about single agents, *PAL* offers a much richer world of social scenarios where agents communicate, like games or dialogues. These usually constrain what can be said. Life is civilized just because we do not 'tell it like it is'. Even so, what can a group achieve by optimal communication? Consider two agents in a model ***M***, *s*. They can tell each other what they know, cutting ***M*** down to smaller parts. If they are cooperative, what is the best information they can give via successive updates – and what does the final collective information state look like?

Example The best agents can do by internal communication.
What is the best that can be achieved in the following model (van Benthem 2002B)?



Geometrical intuition suggests that this must be:



This is correct, and we leave a description of the moves to the reader. ■

In Chapters 12, 15, we will look at such scenarios in more detail, as they model game solution or social deliberation. In particular, in the finite case, agents will normally reach a minimal submodel where further true announcements have no effect. In a sense that can be made precise, they have then converted their *factual distributed knowledge* into *common knowledge*. With two agents, it can be shown that just 2 announcements suffice.

---

[50] In the converse direction, it may be true that after the exam I know that I have failed, but it still need not be the case that I know right now that after the exam I have failed.

[51] It does have intriguing epistemic features of its own, like forgetting or avoiding information.

Example Optimal communication in two existentialist steps.

A two-step solution to the preceding example is the following conversation:

1 sighs: "I don't know what the real world is"

2 sighs: "I don't know either"

It does not matter if you forget details, because it also works in the opposite order. ■

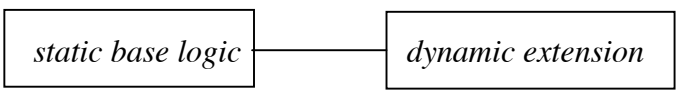
3.5 The dynamic methodology

Having seen some of the interesting issues that *PAL* brings within the compass of logic, let us now try to tease out the general methodology behind public announcement logic.

Compositional analysis Together, the *PAL* axioms analyze the effects of new information compositionally, breaking down the 'post-conditions' behind the dynamic modalities. The key principle here is the recursion axiom for knowledge gain after informational events, and it provides a model of analysis for all further dynamic phenomena that we will study in this book: from belief revision to preference change or group formation.

Static reduction and complexity The recursive character of the *PAL* axioms drives an effective reduction of all assertions with dynamic action modalities to static epistemic statements. This allowed us to establish completeness via a completeness theorem for the static language, while also, a decidable base logic gets a *decidable* dynamic extension. In particular, public announcement logic is decidable because basic epistemic logic is. Thus, in a sense, the dynamics comes for free: we just see more by making it explicit. [52]

Fitting a dynamic superstructure to a static base The general picture here is as follows, and it will recur in this book. First, one chooses a static language and matching models that represent information states for groups of agents, the snap-shots or stills for the dynamic process we are after. Then we analyze the relevant informative events as updates changing these models. Next, these events are defined explicitly in a dynamic extension of the base language, that can also state effects of events by propositions true after their occurrence. This adds a dynamic superstructure over an existing system with a two-tier set-up:



[52] In a little while, we will see also see reasons for going beyond this reductionist approach.

At the static level, one gets a complete axiom system for whatever models one has chosen. On top of that, one seeks dynamic *recursion axioms* that describe static effects of events. In cases where this works, every formula is equivalent to a static one – and the above benefits follow. This design of dynamic epistemic logics is modular, and independent from specific properties of static models. In particular, the *PAL* axioms make no assumptions about accessibility relations. Hence our completeness theorem holds just as well on arbitrary models for the minimal logic *K*, serving as some minimal logic of *belief*. [53] [54]

Pre-encoding and language design Our recursive mechanism needs a static base language with operators that can do conditionalization, as in the clause '*K(P →*' of the Knowledge Gain Axiom. We call this *pre-encoding*: a static model already has all information about what happens after informative events take place. Pre-encoding is ubiquitous, witness the key role of conditionals in belief revision (Chapter 7). If the base language lacks this power, we may have to *redesign* it to carry the weight of its natural dynamics:

Extending PAL with common knowledge Public announcement logic was designed to reason about what people tell each other, and it is quite successful in that. Nevertheless, its basic axioms have no special interaction principles relating different agents. The social character of *PAL* only shows in formulas with iterated epistemic modalities. But what about the acme of that: common knowledge? So far, we have not analyzed this at all.

---

[53] Indeed, this is how some core texts on *PAL* and general *DEL* set up things from the start.

[54] *Preserving frame conditions*. Some interplay between statics and dynamics can occur after all. Suppose we have a static special condition on models, say, accessibility is an equivalence relation. Then a constraint arises on the update mechanism: it should *preserve these conditions*. There is no guarantee here, and one must look case by case. Here is a general fact about *PAL*, given how its update rule works: *PAL* respects any condition on relations that is preserved under submodels. This includes all frame conditions definable by *universal first-order sentences*. But *PAL* update need not preserve existential properties, like every world having a successor. Things are more fragile with the *DEL* product update of Chapter 4, that outputs submodels of direct products of epistemic models. In that case, the only first-order frame conditions automatically preserved are *universal Horn sentences*. Reflexivity, symmetry, and transitivity are of that form, but a non-Horn universal frame condition like linearity of accessibility is not, and hence it may be lost in update.

As it turns out, there just is no *PAL* reduction axiom for $[!P]C_G\varphi$. This is strange, as one main purpose of a public announcement was producing the latter group knowledge. [55] The solution turns out to be the following redesign (van Benthem 1999B). We need to enrich the standard language of epistemic logic in Chapter 2 with a new operator:

Definition Conditional common knowledge.
Conditional common knowledge $C_G^P\varphi$ says that φ is true in all worlds reachable from the current one by a finite path of accessibilities running only through worlds satisfying P. ■

Plain $C_G\varphi$ is just $C_G^T\varphi$. Van Benthem, van Eijck & Kooi 2006 show that $C_G^P\varphi$ cannot be defined in epistemic logic with just common knowledge. The new operator is natural. It is bisimulation-invariant, and also, standard completeness proofs (cf. Fagin, Halpern, Moses & Vardi 1995) carry over to the natural axioms for $C_G^P\varphi$. On top of these static principles, here is a valid recursion axiom for common knowledge in *PAL*:

Fact The following equivalence is valid: $[!P]C_G\varphi \leftrightarrow (P \rightarrow C_G^P[!P]\varphi)$.

Proof Just check with a graphical picture like we did for the Knowledge Gain Axiom. ■

But we have a richer base language now, so we need recursion axioms that work for $C_G^P\varphi$ itself, not just $C_G\varphi$. The next and final axiom shows that the hierarchy stops here:

Theorem *PAL* with conditional common knowledge is axiomatized completely
by adding the valid reduction law $[!P]C_G^\varphi\psi \leftrightarrow (P \rightarrow C_G^{P \wedge [!P]\varphi}[!P]\psi)$.

Example Atomic announcements produce common knowledge.
Indeed, $[!q]C_Gq \leftrightarrow (q \rightarrow C_G^q[!q]q) \leftrightarrow (q \rightarrow C_G^q T) \leftrightarrow (q \rightarrow T) \leftrightarrow T$. ■

By now, we have teased out most general features out of our dynamic pilot system.

We conclude with a possible objection, and a further twist to our methodology.

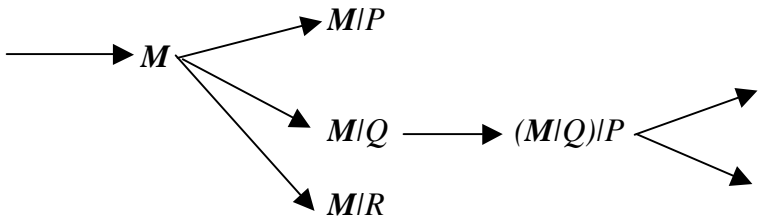
Reduction means redundancy? If *PAL* reduces every dynamic formula to an equivalent static one, does not it shoot itself in the foot, since the dynamics is essentially redundant? Here are a few thoughts. First, the *PAL* axioms are of intrinsic interest, raising many issues that no one had thought of before. The fact that they drive a reduction is secondary – and

[55] Baltag, Moss & Solecki 1998 do axiomatize *PAL* with C_G, using special inference rules.

reductions in science often say much less anyway than it seems at first sight. The heart of the information dynamics is the recursive reduction process *itself*. [56] And its dynamic language brings out phenomena that lay hidden in the epistemic base. But I even like the brute reduction, since it gives substance to my earlier slogan that 'logic can be more than it is'. Existing systems of logic have a much broader sweep than one might think: and this is seen by adding a dynamic superstructure, and then eliciting insights about actions.

Protocols and relaxing the reduction program Even so, there are better reasons for relaxing the orthodox version of the dynamics program, and we shall do so in Chapter 11. Here is the consideration (cf. van Benthem, Gerbrandy, Hoshi & Pacuit 2007). Single steps of communication or observation, as described by *PAL*, make best sense inside longer-term scenarios of conversation or learning. And such scenarios may *constrain* the available sequences of assertions or observations. Not everything that is true can be said or observed.

Now *PAL* itself has no such constraints. It lives in a 'Supermodel' of all epistemic models related by all possible updates, where all histories of true assertions are possible: [57]



Once we accept constraints on histories in this full information process (cf. the *protocols* of Chapter 11), a simple but characteristic axiom of *PAL* so far must be abandoned, viz. the equivalence between the truth of a proposition and its 'executability' (we use the dual existential dynamic modality now, since it fits the current issue better):

$$\varphi \leftrightarrow \langle !\varphi \rangle T$$

Only the implication $\langle !\varphi \rangle T \rightarrow \varphi$ remains valid: announcability implies truth. This change suggests a fundamental conceptual distinction (van Benthem 2009E, Hoshi 2009; cf.

[56] The arithmetic of + is not 'just that of the successor function': the recursion doing the work is crucial. Extensions of theories by definitions are often highly informative, not in a semantic sense, but in a fine-grained intensional sense of information (cf. the syntactic approach of Chapter 5).

[57] Van Benthem 2003D has representation results for the *PAL* language that use only small 'corners' of this huge class, consisting of some initial model $\boldsymbol{M}$ plus a set of further submodels.

Chapter 13) between *epistemic information* about facts in the world and what others know, and *procedural information* about the total process creating the epistemic information. *PAL* only describes the former kind of information, not the latter. While *PAL* with protocols still has recursion axioms for knowledge, they no longer reduce formulas to exclusively epistemic ones. The procedural information that they contain may be essential.

3.6 Conclusion

The logic *PAL* of public announcement, or public observation, is a proof-of-concept: dynamic logics dealing with information flow look just like systems that we know, and can be developed maintaining the same technical standards. But on top of that, these systems raise a host of new issues, while suggesting a larger logical program for describing the dynamics of rational agency with much more sophisticated informational evens than public announcements. We will turn to these in the following chapters.

At first sight so demure and almost trivial, *PAL* generates many interesting questions and open problems. We will now discuss a number of these, though what follows can be skipped without loss of continuity. This is the only chapter in the book where we discuss so many technical issues, and we apologize for its inordinate length. The reason is that *PAL* is a nice pilot system where things are simple to state. But in principle, every point raised here also makes general sense for our later logics of belief, preference, or games.

3.7 Model theory of learning

For a start, we continue with the earlier semantic analysis of public announcement.

Bisimulation invariance Recall the notion of *bisimulation* from Chapter 2, the basic semantic invariance for the epistemic language. *PAL* still fits in this framework:

Fact All formulas of *PAL* are invariant for bisimulation.

Proof This may be shown via the earlier reduction to purely epistemic formulas, that were invariant for bisimulations. But a more informative proof suggesting generalizations goes inductively via an interesting property of update, viewed as an operation O on models. We say that such an operation O *respects bisimulation* if, whenever two models $\mathbf{M}, s$ and $\mathbf{N}, t$ are bisimilar, then so are their values $O(\mathbf{M}, s)$ and $O(\mathbf{N}, t)$:

Fact Public announcement update respects bisimulation.

Proof Let $\leftrightarrows$ be a bisimulation between $\boldsymbol{M}, s$ and $\boldsymbol{N}, t$. Consider the submodels $\boldsymbol{M}|\varphi, s$ and $\boldsymbol{N}|\varphi, t$ after public update with φ. The point is that the *restriction* of $\leftrightarrows$ to these is still a bisimulation. Here is a proof of the zigzag clause. Suppose that some world w in $\boldsymbol{M}|\varphi$ with a $\leftrightarrows$-matching world w' in $\boldsymbol{N}|\varphi$ has an $\sim_i$-successor v in $\boldsymbol{M}|\varphi$. The original bisimulation then gives an $\sim_i$-successor v' for w' *in* $\boldsymbol{N}$ for which $v \leftrightarrows v'$. Now the world v satisfies φ in $\boldsymbol{M}$, and hence, by the Invariance Lemma for the bisimulation $\leftrightarrows$, the matching v' satisfies φ as well. But that means that v' is inside the updated model $\boldsymbol{N}|\varphi$, as required. ■

Many other update mechanisms respect bisimulations, including the one in Chapter 4. [58]

Persistence under update We have seen that not all public events $!\varphi$ result in (common) knowledge that φ. Purely factual formulas φ have this property, and so do some epistemic ones, like Kp. Even $\neg Kp$ is preserved (by a simple argument involving reflexivity), but the Moore formula $p \wedge \neg Kp$ was a counter-example. This raises a general, and as yet unsolved, issue of persistence under update, sometimes called the Learning Problem:

Open Problem Characterize the syntax of the epistemic formulas φ that remain true in a model $\boldsymbol{M}, s$ when the $\neg\varphi$-worlds are eliminated from the model $\boldsymbol{M}$.

Here is a relevant result from modal logic:

Theorem The epistemic formulas in the language without common knowledge that are preserved under submodels are precisely those definable using literals $p, \neg p$, conjunction, disjunction, and K-operators.

Proof Compare the universal formulas in first-order logic, that are just those preserved under submodels. Andréka, van Benthem & Németi 1998 have the technical proof. ■

A conjecture for the full language might add arbitrary formulas $C_G\varphi$ as persistent forms. Van Ditmarsch and Kooi 2006 discuss the case with dynamic modalities. The problem may be hard since lifting first-order model theory to non-first-order modal fixed-point logics (for the modality C_G) seems non-trivial, even on a universe of finite models.

---

[58] Cf. van Benthem 1996 and Hollenberg 1998 on analogous issues in process algebra.

In any case, what we need in the *PAL* setting is not full submodel preservation, but rather preservation under 'self-defined submodels'. When we *restrict* a model to those of its worlds that satisfy φ, then φ should hold in that model, or in terms of an elegant validity:

$$\varphi \rightarrow (\varphi)^{\varphi}$$

Open Problem Which epistemic or first-order formulas imply their self-relativization?

Digression: enforcing persistence? Some people find non-persistence a mere side effect of infelicitous wording. E.g., when ***A*** said "*p*, but you don't know it", she should just have said "*p*", keeping her mouth shut about my mental state. Can we rephrase any message to make the non-persistence go away? An epistemic assertion φ defines a set of worlds in the current model ***M***. Can we always find an equivalent persistent definition? This would be easy if each world has a simple unique factual description, like hands in card games. But there is a more general method that works in more epistemic settings, at least locally:

Fact In each finite *S5*-model, every public announcement has a persistent equivalent. [59]

Of course, the recipe for rephrasing your assertions in a persistent manner is ugly, and not recommended! Moreover, it is local to one model, and does not work uniformly. [60]

---

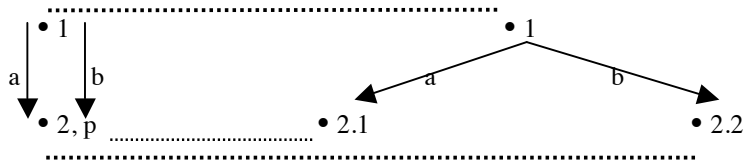
[59] *Proof.* Assume that models ***M*** are bisimulation-contracted (cf. Chapter 2) and connected, with no isolated accessibility zones. Now *j* publicly announces φ, going to the sub-model $\mathbf{M}|\varphi$ with domain $\varphi^* = \{x \in \mathbf{M} \mid \mathbf{M}, x \models \varphi\}$. If this is still ***M*** itself, the persistent announcement "True" works. Now suppose φ^* is not all of ***M***. Our equivalent persistent assertion has two disjuncts: $\Delta \vee \Sigma$. Using the proof of the State Definition Lemma in Chapter 2, Δ is an epistemic definition for φ^* in ***M*** describing each world in it up to bisimulation, and taking the disjunction. Now for Σ. Using the same proof, take a formula describing the model $\mathbf{M}|\varphi$ up to bisimulation, with common knowledge over an epistemic formula describing a pattern of zones and links in $\mathbf{M}|\varphi$ (but no specific world description added). Here is how the new statement works. First, $\Delta \vee \Sigma$ is common knowledge in $\mathbf{M}|\varphi$, because Σ is. But it also picks out the right worlds in ***M***. Clearly, any world in φ^* satisfies its own disjunct of Δ. Conversely, let world *t* in ***M*** satisfy $\Delta \vee \Sigma$. If it satisfies some disjunct of Δ, then *t* is in φ^* by bisimulation-minimality of ***M***. Otherwise, $\mathbf{M}, t$ satisfies Σ. But then by connectedness, every world in ***M*** satisfies Σ, and by the construction of Σ, there is a bisimulation between ***M*** and $\mathbf{M}|\varphi$. But this contradicts the fact that the update was proper. ■

[60] Van Benthem 2006C shows that no uniform solution is possible over all models.

The test of language extensions Finally, how sensitive is this theory to generalization? Sometimes we want to express more than what is available in the basic epistemic language – and this poses an interesting challenge to the scope of *PAL*. An example in Chapter 2 was the notion D_G of distributed group knowledge, that crosses a clear semantic threshold:

Fact The modality $D_G\varphi$ for *intersection* of relations [61] is not invariant for bisimulation.

Proof The following models have $D_{\{a, b\}}\neg p$ true in world *1* on the right, but not on the left:



There is an obvious bisimulation with respect to R_a, R_b – but zigzag fails for $R_a \cap R_b$. ■

Does *PAL* extend to this non-bimulation invariant generalization? The answer is positive:

Theorem *PAL* with distributed knowledge is axiomatized by the earlier principles plus the recursion axiom $[!P]D_G\varphi \leftrightarrow (P \rightarrow D_G[!P]\varphi)$.

Similar observations hold for other extensions of the epistemic base language. For instance, later on, we will have occasion to also use a global *universal modality* $U\varphi$ saying that φ is true in all worlds of the current model, epistemically accessible or not:

Theorem *PAL* with the universal modality is axiomatized by the earlier principles plus the recursion axiom $[!P]U\varphi \leftrightarrow (P \rightarrow U(P \rightarrow [!P]\varphi))$.

So, the methodology proposed here extends well beyond the basic epistemic language.

3.8 Abstract postulates on update and frame correspondence

Let us now step back from our update mechanism. *PAL* revolves around one concrete way of taking incoming hard information. But why should we accept this? What if we reversed the perspective, and asked ourselves which general postulates look plausible a priori for hard information update, and then see *which operations on models* would validate it?

One answer is via a frame correspondence argument (van Benthem 2007B), switching perspectives by taking *PAL* axioms as abstract postulates. We consider only a simple version.

---

[61] On the other hand, adding the modality D_G does keep the epistemic logic decidable.

Let abstract model-changing operations $♥p$ take pointed epistemic models $\boldsymbol{M}, s$ with a set of worlds p in $\boldsymbol{M}$ to new epistemic models $\boldsymbol{M}♥p, s$ – where the domain of worlds remains the same. For still more simplicity in what follows, we also assume that $s \in p$. Now, by *link elimination*, we mean the operation on models that changes accessibility by cutting all epistemic links between p-worlds and $\neg p$-ones. [62]

Theorem Link elimination is the only model-changing operation that satisfies the equivalence $[♥p]Kq \leftrightarrow (p \rightarrow K(p \rightarrow [♥p]q))$ for all propositions q.

Proof We identify propositions with sets of worlds. Start from any pointed model $\boldsymbol{M}, s$. From left to right, take q to be the set of worlds $\sim$-accessible from s in the model $\boldsymbol{M}♥p$. This makes $[♥p]Kq$ true at s. Then the right-hand side says that $K(p \rightarrow [♥p]q)$ holds, i.e., all p-worlds that are $\sim$-accessible from s in $\boldsymbol{M}$ are in q. Thus, the operation $♥p$ preserves all existing $\sim$-arrows from s in $\boldsymbol{M}$ that went into the set p. In the converse direction, let q be the set of all p-worlds that are $\sim$-accessible from s in $\boldsymbol{M}$. This makes $K(p \rightarrow [♥p]q)$ true at s, and hence $[♥p]Kq$ must be true. But the latter says that all worlds that are $\sim$-accessible from s in $\boldsymbol{M}♥p$ are in q: that is, they can only have been earlier $\sim$-accessible worlds in $\boldsymbol{M}$. The two inclusions describe precisely the link-cutting version of epistemic update. ■

This is just one version in a sequence. Here is a next step. Assume now that also the domain of models may change from $\boldsymbol{M}$ to $\boldsymbol{M}♥p$. To zoom in on *PAL*-style eliminative updates, we also need an *existential modality* $E\varphi$ (dual to the universal modality $U\varphi$):

Theorem Eliminative update is the only model-changing operation that satisfies the following three principles: (a) $\langle !p\rangle T \leftrightarrow p$, (b) $\langle ♥p\rangle Eq \leftrightarrow p \wedge E\langle ♥p\rangle q$, and (c) $[♥p]Kq \leftrightarrow (p \rightarrow K(p \rightarrow [♥p]q))$.

Proof sketch (a) $\langle !p\rangle T \leftrightarrow p$ makes sure that inside a given model $\boldsymbol{M}$, the only worlds surviving into $\boldsymbol{M}♥p$ are those in the set denoted by p. Next, axiom (b) $\langle ♥p\rangle Eq \leftrightarrow p \wedge E\langle ♥p\rangle q$ holds for eliminative update, being dual to the earlier one for U. What it enforces

[62] What follows is a bit informal. More precise formulations would use the earlier Supermodel perspective, now with smaller families of epistemic models related by various transformations. We hope that the reader can supply formal details after a first pass through the correspondence proof. The general theme of logics for model transformations gets more interesting with the system *DEL* in Chapter 4, since we then also have constructions that can increase model size

in our abstract setting is that the domain of $\boldsymbol{M}^{\heartsuit}p$ contain no objects beyond the set p in $\boldsymbol{M}$. Finally, the above axiom (c) for knowledge ensures once more that the epistemic relations are the same in $\boldsymbol{M}$ and $\boldsymbol{M}^{\heartsuit}p$, so that our update operation really takes a *submodel*. ■

These results may suffice to show how an old modal technique acquires a new meaning with epistemic update, reversing the earlier direction of conceptual analysis.

3.9 Update versus dynamic inference

This chapter is about the information dynamics of observation, ignoring the dynamics of inference over richer syntactic information states, that will be touched upon in Chapter 5. Even so, *PAL* also suggests a local dynamic notion of inference summarizing the effects of successive updates (Veltman 1996, van Benthem 1996) that is worth a closer look:

Definition Dynamic inference.
Premises $P_1, ..., P_k$ *dynamically imply* conclusion φ if after updating any information state with public announcements of the successive premises, all worlds in the end state satisfy φ. Stated in *PAL* terms, the following implication must be valid: $[!P_1] \ldots [!P_k]C\varphi$. ■

This notion behaves quite differently from standard logic in its premise management:

Fact All structural rules for classical consequence fail.

Proof (a) Order of announcement matters. Conclusions from A, B need not be the same as from B, A: witness the Moore-style sequences $\neg Kp \,;\, p$ (consistent) versus $p \,;\, \neg Kp$ (inconsistent). (b) Multiplicity of occurrence matters, and Contraction fails. $\neg Kp \wedge p$ has different update effects from the inconsistent $(\neg Kp \wedge p); (\neg Kp \wedge p)$. (c) Non-monotonicity. Adding premises can disturb conclusions: $\neg Kp$ implies $\neg Kp$, but $\neg Kp, p$ does not imply $\neg Kp$. (d) Similar dynamic counterexamples work for the Cut Rule. ■

These failures all reflect the essential dynamic character of getting epistemic information. Still, here are three modified structural rules that still hold (van Benthem 1996):

Fact The following three structural rules are valid for dynamic inference:

Left Monotonicity	$X \Rightarrow A$ implies $B, X \Rightarrow A$
Cautious Monotonicity	$X \Rightarrow A$ and $X, Y \Rightarrow B$ imply $X, A, Y \Rightarrow B$
Left Cut	$X \Rightarrow A$ and $X, A, Y \Rightarrow B$ imply $X, Y \Rightarrow B$

Theorem The structural properties of dynamic inference are axiomatized completely by Left Monotonicity, Cautious Monotonicity, and Left Cut. [63]

Thus, *PAL* suggests new notions of consequence in addition to classical logic, and as such, it may be profitably compared to existing *substructural logics*. Van Benthem 2008D has a comparison with the program of Logical Pluralism, a topic taken up in Chapter 13.

3.10 The complexity balance

The third main theme in Chapter 2 was computational complexity of logical tasks.

Theorem Validity in *PAL* is decidable.

Proof This may be shown by the same effective reduction that gave us the completeness, since the underlying epistemic base logic is decidable. ■

This does not settle the computational *complexity* – as translation via the axioms may increase the length of formulas exponentially. [64] Lutz 2006 uses a non-meaning-preserving but polynomial-time *SAT*-reduction to show that *PAL* sides with epistemic logic:

Theorem The complexity of satisfiability in *PAL* is *Pspace-complete*.

As for the other two important items in the complexity profile of our system, we have the following results (van Benthem, van Eijck & Kooi 2006):

Theorem In *PAL*, both model checking and model comparison take *Ptime*.

Both these tasks have interesting two-player game versions, extending those for epistemic logic in Chapter 2. In all then, from a technical point of view, *PAL* presents about the same good balance between expressive power and complexity as its static base logic *EL*. [65]

---

[63] A proof is in van Benthem 2003D. One uses an abstract representation from van Benthem 1996, plus a bisimulation taking any finite tree for modal logic to a model where (a) worlds w go to a family of epistemic models $\boldsymbol{M}_w$, (b) basic actions a go to epistemic announcements $!(\varphi_a)$.

[64] Putting this more positively, dynamic modalities provide very *succinct notation*.

[65] Still, to really understand the complexity profile of a dynamic logic like this, we may have to add *new basic tasks* to the profile, having to do with the complexity of communication.

3.11 The long run: programs and repeated announcements

As we saw with the Muddy Children, to make *PAL* into a logic of conversation or longer-term informational processes, we need program structure for complex assertions and plans. An obvious extension uses *propositional dynamic logic* (*PDL*), an important system for describing abstract sequential actions, whose details we take for granted in this book (cf. also Chapters 4, 10). The reader may consult Blackburn, de Rijke & Venema 2000, Harel, Kozen & Tiuryn 2000, Bradfield & Stirling 2006, or van Benthem, to appearB.

Program structure in conversation: PAL plus PDL We already noted that that repeated assertions can be contracted to one, given the validity of $[!P][!Q]\varphi \leftrightarrow [!(P \wedge [!P]Q)]\varphi$. But despite this reduction, real conversation is driven by complex instructions. If you want to persuade your dean to give money to your ever-hungry institute, you praise him for making things go well if he looks happy, and criticize his opponents if he looks unhappy. And you apply this treatment until he is in a good mood, and then bring up the money issue. This recipe involves all major constructions of imperative programming. Sequential order of assertions is crucial (asking money first does not work), what you say depends on conditions, and flattery involves iteration: it is applied as long as needed to achieve the desired effect. Likewise, the Muddy Children puzzle involved program constructions of

(a) *sequential composition* ;
(b) *guarded choice* *IF ... THEN... ELSE....*
(c) *guarded iteration* *WHILE... DO...*

PAL plus *PDL* program operations on its actions is like *PAL* in that formulas remain invariant for epistemic bisimulation. Still, adding this long-term perspective crosses a dangerous threshold in terms of the complexity of validity:

Theorem *PAL* with all *PDL* program operations added to the action part of the language is undecidable, and even non-axiomatizable.

Miller and Moss 2005 prove this using Tiling (cf. Chapter 2). We will explain what is going on from a general point of view in the epistemic-temporal setting of Chapter 11.

Iterated announcements in the limit Muddy Children was a limit scenario where we keep making the same assertion as long as it is true. Such iterations create intriguing new phenomena. The muddy children repeat an assertion φ of ignorance until it can no longer be made truly. Their statement is *self-defeating*: when repeated iteratively, it reaches a sub-

model of the initial model where it is false everywhere (ignorance turned into knowledge). The other extreme are *self-fulfilling* φ whose iterated announcement eventually makes them common knowledge. The latter occur in solution procedures for games, announcing 'rationality' for all players (cf. Chapters 10, 15 for such *PAL* views of games).

Technically, iterated announcement of φ starting in an initial model $\boldsymbol{M}$ always reaches a *fixed point*. [66] This generalizes our earlier discussion of epistemic formulas φ whose one-step announcement led to truth of $C_G\varphi$ (say, factual assertions), or $C_G\neg\varphi$ (Moore-type formulas). The logical system behind these phenomena is an *epistemic* μ*–calculus* with operators for smallest and greatest fixed-points (Bradfield & Stirling 2006). Chapter 15 has details, and states mathematical properties of *PAL* with iteration in this setting.

3.12 Further directions and open problems

We end with some topics in the aftermath of this chapter that invite further investigation – ranging from practical modeling to mathematical theory.

Art of modeling Fitting epistemic logic and *PAL* to specific applications involves a choice of models to work with. In this book, this art of modeling will not be studied in detail, even though it is crucial to the reach of a framework. It raises general issues that have not yet been studied systematically in our logics. For instance, with public announcements in conversation, there is the issue of *assertoric force* of saying that φ. Normal cooperative speakers only utter statements that they know to be true, generating an event $!(K\varphi)$. They may also convey that they think the addressee does not know that φ. Such preconditions can be dealt with in the dynamic-epistemic logic of Chapter 4 and the protocols of Chapter 11. Weaker preconditions involving beliefs occur in Chapter 7, and others are discussed in our analysis of questions (Chapter 6). But there are many further features. For instance, there is an issue of language choice and *language change* in setting up the range of epistemic possibilities. Speech acts may make new proposition letters relevant, changing the representation space. When these phenomena are systematic enough, they may themselves enter the dynamic logic: the product update mechanism of Chapter 4 is one step toward a systematic view of model construction.

---

[66] One defines repeated announcement in infinite models by taking *intersections* at limit ordinals. In each model, each epistemic formula is then either self-defeating or self-fulfilling.

From analyzing to planning *PAL* has been used so far to analyze given assertions, but it can also help plan assertions meeting specifications. Here is an example from a Russian mathematical Olympiad (cf. van Ditmarsch 2003): "*7* cards are given to persons: *A* gets *3*, *B* gets *3*, C gets *1*. How should *A, B* communicate publicly, in hearing of *C*, to find out the distribution of the cards while *C* does not?" Solutions depend on the number of cards. More generally, how can a subgroup communicate, keeping the rest in the dark? Normally, this needs 'hiding' (cf. Chapter 4), but some tasks can be done in public (cf. van Ditmarsch & Kooi 2008 on reachability from one model to another via announcements). A relevant analogy here comes from program logics with *correctness assertions*

$$\varphi \rightarrow [!P]\psi$$

saying that, if precondition φ holds, then public observation of P always leads to a model where postcondition ψ holds (cf. van Benthem 1996). Given $!P$, one can analyze its pre- and postconditions. In fact, $[!P]\psi$ defines the 'weakest precondition' for $!P$ to produce effect ψ. Or given $!P$ plus precondition φ, we can look for the 'strongest postcondition' ψ. E.g., with $\varphi = True$, and p atomic, it is easy to see that the strongest postcondition is common knowledge of p. [67] This is program analysis. But there is also *program synthesis*. Given precondition φ and postcondition ψ (say, specifying which agents are to learn what) we can look for an assertion $!P$ guaranteeing that transition.

Agents and diversity We have seen how *PAL* makes idealizing assumptions about agents' powers of memory and observation. Liu 2005, 2008 discuss *diversity* in powers, and the need for *PAL* and other dynamic-epistemic logics to model agents with different powers interacting successfully. For this, we need to parametrize update rules (Chapter 4 gives an option). Indeed, *PAL* does not have an explicit notion of *agent* – and maybe it should. [68]

The secret of PAL: relativization Here is the ultimate technical explanation for the *PAL* axioms (van Benthem 1999B). Updating with P is *semantic relativization* of a model $\boldsymbol{M}, s$ to a definable submodel $\boldsymbol{M}|P, s$. Now standard logic has a well-known duality here. One

---

[67] In general, though, epistemic post-conditions need not be definable in *PAL*: cf. Chapter 11.

[68] Agents *themselves* might change by learning. But then our recursion axiom $[!P]K_i\varphi \leftrightarrow (P \rightarrow K_i(P \rightarrow [!P]\varphi))$ is too simplistic. Agent i before update changes into agent $i+!P$, and we must modify the axiom – perhaps to an equivalence more like $[!P]K_{i+!P}\varphi \leftrightarrow (P \rightarrow K_i(P \rightarrow [!P]\varphi))$.

can either evaluate epistemic formulas in the new model $\boldsymbol{M}|P$, or translate them into formulas in the old model $\boldsymbol{M}, s$ by means of *syntactic relativization*:

Relativization Lemma $\quad \boldsymbol{M}|P, s \models \varphi$ iff $\boldsymbol{M}, s \models (\varphi)^P$

The recursion axioms of *PAL* are just the inductive clauses in syntactic relativization. And our 'pre-encoding' becomes the issue if the base language is closed under relativization. *EL* with common knowledge was not, [69] and conditional common knowledge solved this. In this light, *PAL* is a logic of relativization, related to logics of other basic operations such as *substitution,* that can be performed both semantically and syntactically. [70]

Schematic validity Next comes an open problem that I find slightly annoying. Unlike with most logical systems, the axioms of *PAL* are not closed under substitution of arbitrary formulas for atoms. For instance, the base axiom fails in the form $[!P]Kq \leftrightarrow (P \rightarrow Kq)$. With $P = q$, $[!q]Kq$ is always true, while the right-hand side $q \rightarrow Kq$ is not valid. Still, except for the atomic case, all *PAL* axioms are *schematically valid*: each substitution of formulas for Greek letters preserves validity. Even in the base, we have schematic validity of $\langle!\varphi\rangle T \leftrightarrow \varphi$. Another nice example was our earlier law for repeated announcements, showing that there are interesting schematic validities that are not immediately apparent from the *PAL* axioms. The schematically valid principles are the truly general algebraic laws of public announcement. But it is not clear how to describe them well. For instance, our completeness proof does not show that schematically valid laws are derivable *as such*: we have only shown that each of their concrete instances is derivable.

Open Problem Are the schematic validities of *PAL* axiomatizable? [71]

Model theory of iteration, learning, and fixed-points There are many open problems in dynamic logics and μ–calculi extending *PAL* with program structure. Just ask yourself this: how do updates for related formulas compare? Suppose that φ implies ψ. Will the iterated $!\varphi$–sequence always go faster than that for $!\psi$? Are their limits included? Chapter 15 shows that neither needs to be the case, and asks further questions (cf. also van Benthem 2007F).

---

[69] A related non-relativizing system is propositional dynamic logic *without tests*. In contrast, the complete language of propositional dynamic logic with tests is closed under relativization.

[70] What is the complete logic of relativization $(\varphi)^A$ in first-order logic? Van Benthem 1999B notes that the schematically valid iteration law of *PAL* here becomes *Associativity* $((A)^B)^C \leftrightarrow_A ((B)^C)$.

[71] An answer is not obvious since schematic validity quantifies over all formulas of the language.

But mysteries abound. For instance, the *PDL* extension of *PAL* looks simple, but Baltag & Venema have shown that the formula $<(!P)^*>C_G q$ is not even definable in the modal μ–calculus, as it lacks the finite model property. Intuitively, formulas $[(!P)^*]\varphi$ are still definable with greatest fixed-point operators of some sort, but we lack a good view of the best formalisms. Finally, to return to a concrete issue, recall the Learning Problem of determining just which *PAL*-formulas φ become common knowledge upon announcement: i.e., $[!\varphi]C_G\varphi$ is valid. Generalized issues of this sort arise with iteration. Say, when is a formula φ self-confirming or self-refuting in the announcement limit?

3.13 Literature

In this final section, I give a few key references as markers to the field, and the same sort of bibliography will conclude later chapters. My aim is not a survey of contributions, nor an official record of credits, though I try to be fair. Update of semantic information as reduction in a range of possible situations occurs all across science (think of probabilistic conditioning), and also in the world of common sense. There have been attempts to patent this idea to particular authors, but these are misguided. Systematic logics of update have been proposed in the semantics and pragmatics of natural language (Stalnaker 1978, Veltman 1996). The *PAL* style logic of model-changing announcements is due to Plaza 1989, though this paper was unknown and had no direct influence. It was rediscovered independently by students at ILLC Amsterdam. Gerbrandy & Groeneveld 1997 documents a first stage, Gerbrandy 1999A goes further. Baltag, Moss & Solecki 1998 axiomatize *PAL* with common knowledge. Van Benthem 2006C adds new themes and results that occur in this chapter, obtained since the late 1990s. Van Ditmarsch, van der Hoek & Kooi 2007 is a textbook on general dynamic-epistemic logic, but including much material on *PAL*. Van Benthem, van Eijck & Kooi 2006 has a streamlined version with recursion axioms for common knowledge, and definability results for *PAL* are proved with model-theoretic techniques like Ehrenfeucht games. Miller & Moss 2005 is a sophisticated mathematical study of the complexity of *PAL* with program constructions. The first significant philosophical uses of *PAL* are in Gerbrandy 1999A, B and van Benthem 2004B.