Bisimulations on Planet Kripke

Jelle Gerbrandy

Bisimulations on Planet Kripke

ILLC Dissertation Series DS–1999-01

institute *for* logic, language *and* computation

For further information about ILLC-publications, please contact

Institute for Logic, Language and Computation
Universiteit van Amsterdam
Plantage Muidergracht 24
1018 TV Amsterdam
phone: +31-20-5256090
fax: +31-20-5255101
e-mail: illc@wins.uva.nl

Bisimulations on Planet Kripke

Academisch Proefschrift

ter verkrijging van de graad van doctor aan de
Universiteit van Amsterdam,
op gezag van de Rector Magnificus
prof.dr. J.J.M. Franse
ten overstaan van een door het college voor promoties ingestelde
commissie in het openbaar te verdedigen in de
Aula der Universiteit
op 26 januari 1999 te 15.00 uur

door

Jelle Douwe Gerbrandy

geboren te Amsterdam

Promotor: Prof.dr. J.A.G. Groenendijk
Afdeling Wijsbegeerte
Faculteit der Geesteswetenschappen
Universiteit van Amsterdam
Nieuwe Doelenstraat 15
1012 CP Amsterdam

Copyright © 1998 by Jelle Gerbrandy

Printed by Print Partners Ipskamp

ISBN: 90–5776–019–3

Acknowledgements

These are the people and institutions, and combinations thereof, that I would like to thank, for different reasons having to do with this dissertation as well as with my personal self. In alphabetical order:

Peter Aczel, Giovanna D'Agostino, Chris Albert, Maria Aloni, Alexandru Baltag, Renate Bartsch, Ria Beentjes, Peter Blok, Elsbeth Brouwer, Johan van Benthem, Elena Brosio, the Center for the Study of Language en Information (CSLI) at Stanford University, Marijne van Dam, Paul Dekker, the Dutch Science Organization (NWO), Peter van Emde Boas, Anna Gerbrandy, Gerbrich Gerbrandy, Jeltsje Gerbrandy, Sjoerd Gerbrandy, Anastasia Giannakidou, Robert Goijers, Willem Groeneveld, Joseph Halpern, Paul Harrenstein, Herman Hendriks, Wiebe van der Hoek, Marco Hollenberg, the Institute of Logic, Language and Computation (ILLC), Gerhard Jäger, Dick de Jongh, Gwen Kerdiles, Donald Knuth, Denise Koeleveld, Karen Kwast, Barteld Kooi, Saul Kripke, Leslie Lamport, Jaap Maat, Maarten Marx, Nikos Massios, Anne-Marie Mineur, Judith Moody, Larry Moss, Hans de Nivelle, Anna Pilatova, Marjorie Pigge, Jan Popkema, Robert van Rooy, Hans Rott, Jenny Slatman, Harry Stein, Martin Stokhof, Allard Tamminga, Alfred Tarski, Kees Vermeulen, the University of Amsterdam Frank Veltman, Yde Venema, Bruno Verbeek, Marco Vervoort, Marco de Vries, Piter de Weerd, Reni Webb, Henk Zeevat

Special thanks go to Jeroen Groenendijk, who, as my supervisor, taught me, by example as well as by instruction, the morals of doing research of the kind reported in this book.

Contents

Introduction

In some scientific journals it is customary to publish a list of key words above each article. This is very convenient, because one can see in one glance whether one should read the article or not.

Here are some key words for this dissertation, with a short explanation of each.

Epistemic Logic is the logic of knowledge and belief.

Multi-agent Epistemic Logic is a version of epistemic logic that is concerned with the beliefs of several agents, including the beliefs they have about each other's beliefs.

Dynamic Epistemic Logic is the logic of changes in knowledge or belief. It is the logic of information change.

Non-well-founded Set-theory is a version of set-theory that differs from the more classical set-theory in that there exist sets that are not well-founded. I use this theory to give a semantics for epistemic logic.

I have tried to write the thesis in such a way that all chapters are independent of each other, mostly because this made it much easier for me to write, but also because it means that a reader can read separate chapters and still have some idea what is going on. The only real exception to this modular approach is chapter 5, which does not make sense without the preceding chapters.

Chapter 1 is a short introduction to non-well-founded set theory. I will make a lot of use of this set-theory in this dissertation. It is based on a set of lecture notes from a small informal course I gave to Ph.D. students from the ILLC.

Chapter 2 is a short study of the notion of bisimulation and its role in semantics for epistemic logic. This chapter is mostly technical in character. The main point is to show that several different ways of providing a semantics for epistemic logic are manifestations of one and the same idea, and that bisimulation provides the connection between these different models. This chapter was printed, in almost exactly the same form, as Gerbrandy (1997a).

The next chapter, chapter 3, consists of several parts. I give an introduction to epistemic semantics using non-well-founded sets in section 3.1. The three sections after that are concerned with different notions of 'group knowledge.' It contains a discussion of how to define operators that express common knowledge, distributed knowledge and combined knowledge in possible worlds semantics. The chapter ends with a sketch of how the semantics of knowledge and belief can be extended to a semantics in which one can talk about knowledge about a certain topic. From this chapter, the section on distributed knowledge appeared in more or less the same form as Gerbrandy (1998), while the chapter on common knowledge has some overlap with Gerbrandy (1997b).

Chapter 4 is really the heart of this dissertation. In this chapter Dynamic Epistemic Semantics is defined, which is an extension of epistemic logic in which we can talk about change in information and the results of such changes. The chapter contains a sound and complete axiom system for the logic. The main ideas for the semantics come from Gerbrandy (1994); the semantics was further developed in Gerbrandy and Groeneveld (1997) and Gerbrandy (1997c).

The chapter after that, chapter 5, is concerned with alternatives to the non-well-founded semantics of the chapter 4. I show how one can define the operations of change of dynamic epistemic semantics in other ways, and discuss the work of other authors that have studied logics of information change. Apart from the last section on update semantics, none of the ideas in this chapter have been published before.

Chapter 6 is a study of how the 'common ground,' as it is used in certain theories of dialogue, relates to the notion of 'common knowledge.' It makes use of some of the techniques of the previous chapters to study the relation between changes in the common ground and changes in the knowledge of the agents involved. Most of the contents of this chapter are taken from Gerbrandy (1997b).

In the last chapter, dynamic epistemic semantics is applied to two puzzles: the puzzle of the dirty children, and the surprise examination paradox. The chapter ends with a sketch of a very simple model of dialogue using the techniques of Dynamic Epistemic Semantics. The part on the puzzle of the dirty children appeared in a more condensed form as a part of Gerbrandy (1994) and in Gerbrandy and Groeneveld (1997).

1

Non-well-founded Set Theory

This chapter contains some background material to the theory of sets which is used in most of this dissertation. These pages should provide a reader who already has some basic knowledge of standard set-theory with a rough idea of what non-well-founded set theory is, and with enough information to understand the set-theoretical details of the definitions and proofs in this dissertation. The first part of this chapter is a recapitulation of the first chapter of Aczel (1988). In the second part, section 1.2, the notion of a possibility is introduced. Possibilities will play a central role in the rest of this dissertation; I will use them as models for epistemic logic.

The theory of non-well-founded sets that is the topic of this chapter is set forth by Peter Aczel in his book "Non-Well-Founded Sets" (Aczel, 1988). Other introductions to Aczel's set theory can be found in Barwise and Etchemendy (1988) and in the last chapter of Keith Devlin (1992). The book by Barwise and Moss (1996) explores the theory in more detail and discusses some applications.

Not in these notes are proofs of theorems or detailed technical results. For the proofs, generalizations, and subtleties regarding hypersets, the reader is referred to any of the introductions just mentioned.

1.1 Hypersets

A well-founded set is a set s such that there is no infinite sequence $s_0, s_1, s_2 \ldots$ such that $s_1 \in s_0$, $s_2 \in s_1$, $s_3 \in s_2$, and $s_{i+1} \in s_i$ for all i. The canonical axiomatization of set-theory, the ZFC-axioms,[1] contains an explicit axiom, the

[1] In the following, I will write "ZFC" for the standard Zermelo-Fraenkel-Cantor axioms of set theory without the foundation axiom, and I write "ZFC(AX)" for the axioms of ZFC together

foundation axiom. Together with the other axioms of ZFC it implies that each set is well-founded.

Foundation Axiom (FA)

$$\forall a(\exists x(x \in a) \to \exists x(x \in a \wedge \forall y(y \in x \to \neg(y \in a))))$$

In non-well-founded set-theory, this axiom is replaced by an 'anti-foundation axiom.' In contrast to the foundation axiom, the new axiom implies that there *are* sets containing themselves as elements, functions that have themselves as arguments or as values, etcetera. Before introducing Aczel's axiom of anti-foundation, we first discuss the idea of picturing sets.

1.1.1 Sets and Graphs

One can draw pictures of well-founded sets in the form of graphs: sets are represented by the nodes in the graph, and the edges represent that one set is an element of another. For example, if we identify a natural number n with the set of numbers smaller than n, we can depict the sets 0, 1, 2 and 3 as in figure 1.1.

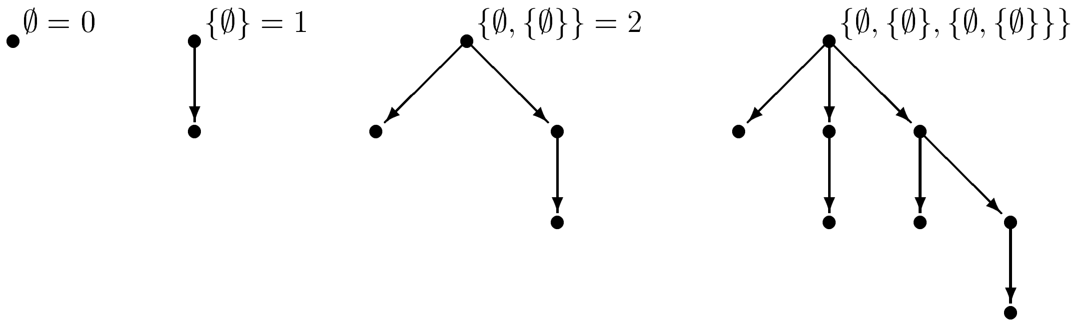


Figure 1.1. Pictures of the sets 0, 1, 2 and 3

To make this idea of graphs picturing sets more precise, I first define what a graph is:

Definition 1.1 (pointed graphs)

- A *graph* $\mathcal{G}$ is a pair $(G, \longrightarrow)$ consisting of a set of nodes G and a relation $\longrightarrow$ over G.
- A *pointed graph* $(\mathcal{G}, n)$ is a graph $\mathcal{G}$ together with a distinguished node n.

---

with the axiom AX.

If $n \longrightarrow n'$, we say that n' is a *successor* of n. A graph is *well-founded* iff it does not contain an infinite path $n \longrightarrow n_1 \longrightarrow n_2 \ldots$ □

As said, graphs can be seen as 'pictures' of sets. To make this precise, we define the notion of decorating a graph with sets by labeling each node in the graph with a set.

Definition 1.2 (decorations and pictures)

- A *decoration* of a graph $\mathcal{G}$ is a function δ that assigns to each node n of $\mathcal{G}$ a set δ_n in such a way that the elements of δ_n are exactly the sets assigned to successors of n, i.e. $\delta_n = \{\delta_{n'} \mid n \longrightarrow n'\}$.
- If δ is a decoration of a pointed graph $(\mathcal{G}, n)$, then $(\mathcal{G}, n)$ is a *picture* of the set δ_n. □

A decoration assigns to each node in a graph a set in such a way that this set contains all and only sets that are assigned to the successors of that node. If a pointed graph $(\mathcal{G}, n)$ has a decoration that assigns the set a to n, then $(\mathcal{G}, n)$ is a picture of a.

Using the axioms of set-theory (but not the axiom of foundation), we can now prove what we said above, namely that each set can be pictured by a (pointed) graph.

Proposition 1.3 In ZFC, it holds that:

$$\text{Every set has a picture} \qquad \square$$

Moreover, it holds that each well-founded graph depicts a well-founded set unambiguously: there is just one single way of decorating a well-founded graph with well-founded sets.

Proposition 1.4 (Mostovski's Collapsing Lemma) It holds in ZFC that:

$$\text{Each well-founded graph is a picture of exactly one set} \qquad \square$$

1.1.2 The Anti-foundation Axiom

The foundation axiom implies that all sets are well-founded. It turns out that the foundation axiom is, in ZFC, provably equivalent with the statement

foundation Only well-founded graphs have decorations.

The foundation axiom says that only well-founded graphs are pictures of sets: no set has a non-well-founded picture. With proposition 1.3, we can conclude that ZFC(FA) implies that sets are exactly those things that can be pictured by well-founded graphs.

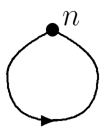


Figure 1.2.

The foundation axiom FA implies that the graph of figure 1.2 is not a picture of any set.

One way of introducing non-well-founded sets in the universe is to replace the foundation axiom with an axiom that implies that *every* graph is a picture of some set.

anti-foundation (AFA_1) Each graph has a decoration.

$\mathsf{ZFC}(\mathsf{AFA}_1)$ implies that sets are exactly those things that can be pictured by some graph, be it a well-founded graph or not.[2] AFA_1 asserts that 'more' sets exist than just the well-founded ones. For example, it implies that the circular picture above has a decoration δ. This decoration assigns to the node n a set δ_n with the property that all elements of δ_n are decorations of successors of n. Since n has only itself as a successor, it holds that $\delta_n = \{\delta_n\}$. Clearly, this set is not well-founded.

Although AFA_1 expresses that non-well-founded sets exist, it gives us only a very weak theory of non-well-founded sets. In particular, it does not tell us very much about the relation of identity between non-well-founded sets. Consider for example the graph of figure 1.3. We know that this graph has a decoration. Suppose that this decoration assigns a set s' to the node n' of this graph.

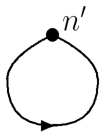


Figure 1.3.

How do we know whether the set s' is the same as the set s which was assigned to n in figure 1.2? The axiom of extensionality, which states that two sets are

[2] This may be confusing: in the definition of a graph, the notion of a set is used in the *definiens*, while at the same time, we make use of graphs to define what sets are. There is nothing mysterious in this, however: this is standard practice in axiomatic set-theory. For example, the pairing axiom says that any two sets can be combined to form a new set that is the pair of the two given sets. Also this axiom uses the concept of a set to define which sets exist. The anti-foundation axiom works in the same way.

equal just in case they have the same elements, is not of much help. The two sets s and s' have only themselves as elements, so we cannot apply it. We need something stronger than extensionality to determine when non-well-founded sets are equal. One way to do that is by taking the hint from proposition 1.4, which says that two well-founded sets are equal if they have a picture in common. We just postulate that all sets are equal just in case they can be pictured by the same graph:[3]

anti-foundation(AFA_2) Each graph has at most one decoration.

This axiom says that each graph is a picture of a unique set: graphs picture sets unambiguously. With AFA_2, we can prove that the sets s and s' we discussed above must be equal. We can decorate the circular graph of figure 1.2 with both s and with s': the circular graph is a picture of s as well as of s'. The axiom AFA_2 tells us that each graph is the picture of at most one set. So s and s' must be the same set.

Some thinking shows that there is only one set which contains itself as its only element. For let t be such a set. Then the graph of figure 1.2 can be decorated with t. So, t is equal to s. The set that has only itself as an element is usually denoted by Ω.

Aczel's anti-foundation axiom is the conjunction of AFA_1 and AFA_2:

Anti-Foundation Axiom (AFA) Every graph is a picture of exactly one set.

Aczel's book contains a proof for the statement that $\mathsf{ZFC}(\mathsf{AFA})$ is consistent just in case ZFC is.

1.1.3 Bisimulation and Equality

In ZFC, it holds that each well-founded graph is a picture of a unique set, and AFA implies that each graph is a picture of a unique set. This does not mean that each set has a unique picture, though. We have already seen that the graphs of figure 1.2 and 1.3 are pictures of the same set. Similarly, the set 2 can be pictured in any of the ways depicted in figure 1.4.

This raises the question when two graphs are pictures of the same set. It turns out that the relation that holds between two graphs when they are pictures of the same set can be characterized by a notion familiar from process algebra and modal logic: that of bisimulation.

Definition 1.5 (bisimulation between graphs)
Let $\mathcal{G}$ and $\mathcal{G}'$ be two graphs.

A relation R between the nodes of $\mathcal{G}$ and the nodes of $\mathcal{G}'$ is a bisimulation iff the following holds for each n in $\mathcal{G}$ and n' in $\mathcal{G}'$ such that nRn':

[3] There are other options: Aczel discusses some variants for this axiom.

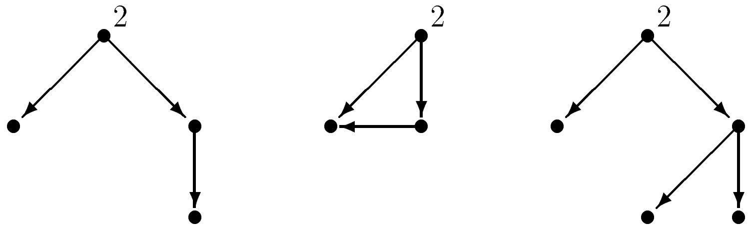


Figure 1.4. Pictures of the set 2

1. For each m such that $n \longrightarrow m$ there is an m' such that $n' \longrightarrow m'$ and mRm'.

2. For each m' such that $n' \longrightarrow m'$ there is an m such that $n \longrightarrow m$ and mRm'.

We say that two pointed graphs $(\mathcal{G}, n)$ and $(\mathcal{G}', n')$ are bisimilar when there is a bisimulation connecting n with n'. □

The following is a theorem in ZFC(AFA):

Proposition 1.6 Let $(\mathcal{G}, n)$ and $(\mathcal{G}', n')$ be two pointed graphs. It holds in ZFC(AFA) that:

$(\mathcal{G}, n)$ and $(\mathcal{G}', n')$ are bisimilar iff they are pictures of the same set. □

Bisimulation is the equivalence relation that connects two graphs exactly when they are pictures of the same set. The notion of bisimulation provides us with a useful tool to prove that two sets are equal in a way that is more direct than by showing that two sets have a picture in common.

Definition 1.7 (bisimulation between sets)
A relation R is a bisimulation on sets iff for all sets s and t, if sRt, then

1. For each $s' \in s$ there is a $t' \in t$ such that $s'Rt'$.

2. For each $t' \in t$ there is an $s' \in s$ such that $s'Rt'$

Two sets s and t are bisimilar iff there is a bisimulation R such that sRt. □

It holds that two sets are equal just in case they are bisimilar:

Proposition 1.8 In ZFC(FA) and in ZFC(AFA), it holds that:

s and t are bisimilar iff $s = t$. □

The fact that two bisimilar sets are equal is the non-well-founded set-theoretical counterpart to the axiom of extensionality. In well-founded set theory, we often show that two sets are equal by showing that they have the same elements. For non-well-founded sets, the axiom of extensionality is not strong enough to prove equality or inequality between sets, and we often use the (stronger) bisimulation result instead.

As is well-known, bisimulation has a straightforward game-theoretical interpretation. This is interesting, because such a game-theoretical interpretation can be an intuitive motivation for identifying bisimilar sets (and bisimilar sets only).

Suppose you and an opponent are playing a game with two sets s and t. Your opponent wants to prove that the two sets are distinct, you defend the claim that they are equal. Two sets are different if one of them contains an element that is not an element of the other set. Your opponent, who makes the first move of the game, chooses an element x from one of the sets, say from set s, and claims that x is not equal to any element of t. You disagree, and to motivate your claim, you choose an element of the other set, say y. You say that y is in fact equal to x. Now you are in the same position in which you started: there are two sets x and y of which you say they are equal, and your opponent says they are not. So you can start playing the game all over again, but this time with x and y.

Your opponent wins the game if at a certain point in the game, you cannot answer his move anymore: in that case, your opponent has found an element of a set x that is not an element of another set y, while you claimed that x and y were equal. In all other cases, you win.

It turns out that if you and your opponent play the game with sets s and t, you have a winning strategy just in case s and t are bisimilar. With the proposition above, this means that you can win the game if and only if s and t are equal. A more formal version of this result is stated in chapter 2.

1.1.4 Equations

An alternative to describing sets in terms of pictures is by using equations. In fact, we have already done this informally, by saying that the set Ω is the only set x that has the property that $x = \{x\}$. This suggests that we can use the equation $x = \{x\}$ to *define* the set Ω.

To formulate the notion of an equation, we need objects that can play the role of 'indeterminates' in the equations. One way to do that is by introducing 'atoms' or 'urelements' in our set theory: objects that are not sets and that have no further set-theoretic structure. We will assume that we have an unlimited supply of urelements, and use these as our indeterminates in the definitions below.

Definition 1.9 (systems of equations)

- For each class of indeterminates $\mathcal{X}$, a *system of equations in the indeterminates* $\mathcal{X}$ is a class of equations that includes for each $x \in \mathcal{X}$ exactly one equation of the form $x = \{x_1, x_2 \ldots\}$, where $\{x_1, x_2 \ldots\}$ is a subset of $\mathcal{X}$. We will write $(x = s_x)_{x \in \mathcal{X}}$ for such a system of equations. □

For example,

$$x = \{x\}$$

is a system of equations in the indeterminates $\{x\}$, and

$$\begin{gathered} x = \{x, y\} \\ y = \emptyset \end{gathered}$$

is a system of equations in the indeterminates $\{x, y\}$. The class of equations of the form $x_\alpha = \{x_{\alpha+1}\}$, where α is any ordinal, is a system of equations in the class of indeterminates $\{x_\alpha \mid \alpha \text{ is an ordinal}\}$.

Definition 1.10 (solutions)
A *solution* of a system of equations $(x = s_x)_{x \in \mathcal{X}}$ is a function δ that assigns to each $x \in \mathcal{X}$ a set δ_x such that $\delta_x = \{\delta_y \mid y \in s_x\}$.
If δ is the solution of a system of equations $(x = s_x)_{x \in \mathcal{X}}$, then the set (or class) $\{\delta_x \mid x \in \mathcal{X}\}$ is called the *solution set* of that system. □

The following result is a direct consequence of the anti-foundation axiom. In fact, it is equivalent to it in ZFC.

Proposition 1.11 (solution lemma)
It holds in ZFC(AFA) that:

Each system of equations has a unique solution. □

The proof of this result is very straightforward once one sees the close connection between graphs and systems of equations.[4] Given a graph, you get a system of equations by taking the nodes of $\mathcal{G}$ as your indeterminates, and letting $x = s_x$ be an equation in the system just in case s_x is the set of successors of x. Vice versa, if we have a system of equations $(x = s_x)_{x \in \mathcal{X}}$, we can take $\mathcal{X}$ as our set of nodes, and let $x \longrightarrow y$ iff $y \in s_x$. In either direction, the solution of a given system of equations will assign the same set to an indeterminate x as in the solution of the corresponding graph.

The solution lemma above can be stated in a more general form, which I will use many times in this dissertation. The *generalised solution lemma* applies to systems of equations in which the sets s_x occurring on the right-hand side of the equations are allowed to contain not only indeterminates as elements, but also sets constructed with or without the use of indeterminates. Under this more general notion, also this

[4]At least, as long as the systems are *sets* of equations. If the system is a proper class of equations, the result is less obvious, and one needs rather heavy set-theoretical machinery to prove that the solution lemma is true in the general case.

$$x = \langle \Omega, \{\emptyset, x\} \rangle$$

is a system of equations. The generalized solution lemma is the result that such systems of equations have unique solutions as well.

1.1.5 Least and Largest Fixed Points

Often, sets and classes of sets are defined as the least fixed point of some operator Φ.

Definition 1.12 (least and largest fixed points)

- A class A is a fixed point of an operator Φ iff $\Phi(A) = A$.
- A class A is the *least fixed point* of an operator Φ iff A is a fixed point of Φ, and if A' is a fixed point of Φ as well, then $A \subseteq A'$.
- A class A is the *largest fixed point* of an operator Φ iff A is a fixed point of Φ, and if A' is a fixed point of Φ as well, then $A' \subseteq A$. □

Not every operator has a least and a largest fixed point. It is well-known that a sufficient condition for the existence of a (unique) least and largest fixed point of an operator is the property of being *set-continuous*.

Definition 1.13 An operator Φ is *(set-)continuous* just in case:
(1) Φ is *monotone*: if $A \subseteq B$, then $\Phi(A) \subseteq \Phi(B)$
(2) Φ is *set-based*: for all $x \in \Phi(A)$, there is a *set* $s \subseteq A$ such that $x \in \Phi(s)$ □

It is a standard result in ZFC that if an operator Φ is set-continuous, the operator has both a least and a largest fixed point. The least fixed points can be characterized in the following way:

Proposition 1.14 (characterization of least fixed points)
Let Φ be a set-continuous operator. Let for each ordinal α, Φ_α be such that:

$$\Phi_\alpha = \bigcup_{\beta<\alpha} \Phi(\Phi_\beta)$$

The least fixed point of Φ is $\bigcup_\alpha \Phi_\alpha$. The sets Φ_α can be defined more vividly as follows:

$$\begin{array}{rcl} \Phi_0 & = & \emptyset \\ \Phi_{\alpha+1} & = & \Phi(\Phi_\alpha) \\ \Phi_\lambda & = & \Phi(\bigcup_{\alpha<\lambda} \Phi_\alpha) \text{ if } \lambda \text{ is a limit ordinal} \end{array}$$

Many interesting sets (or classes) can be, and often are, defined as the least fixed point of some operator. For example, the set of finite ordinals ω is the least fixed point of the following operator Φ_ω:

$$\Phi_\omega(s) = \{\emptyset\} \cup \{\{t\} \cup t \mid t \in s\}$$

I.e. the set of finite ordinals is the smallest class A such that (1) A contains the empty set and (2) if $s \in A$, then $s \cup \{s\} \in A$.

The class of hereditarily finite sets (sets that contain finitely many elements, and for which it holds that each set that is used in its transitive closure contains finitely many elements) can be defined in $\mathsf{ZFC}(\mathsf{AFA})$ as the least fixed point of the operator Φ_{HF}.

$$\Phi_{HF}(s) = \{t \subseteq s \mid t \text{ is a finite set}\}$$

Another example is the definition of the language of propositional logic. Suppose $\mathcal{P}$ is a set of urelements, and so are $\wedge$ and $\neg$, and abbreviate an ordered pair of the from $\langle \neg, s\rangle$ as $\neg s$, and let $s \wedge t$ be an abbreviation of $\langle s, \wedge, t\rangle$. Given this, we can define the set of formulas of propositional logic as the least fixed point of $\Phi_{\mathcal{L}}$:

$$\Phi_{\mathcal{L}}(s) = \mathcal{P} \cup \{\neg\phi \mid \phi \in s\} \cup \{\phi \wedge \psi \mid \phi \in s \text{ and } \psi \in s\}$$

With respect to least fixed points, continuous operators in $\mathsf{ZFC}(\mathsf{AFA})$ behave just as they do in $\mathsf{ZFC}(\mathsf{FA})$:

Proposition 1.15 If Φ is a continuous operator such that for all well-founded A, $\Phi(A)$ is well-founded, then the least fixed point of Φ is well-founded. □

This means that any operator which is defined in the well-founded universe has 'the same' least fixed point in both $\mathsf{ZFC}(\mathsf{AFA})$ and $\mathsf{ZFC}(\mathsf{FA})$. This means that we often do not need to worry about using familiar definitions for inductively defined classes from well-founded set theory in $\mathsf{ZFC}(\mathsf{AFA})$. It also means that if we want to define classes that include non-well-founded sets, inductive definitions may not always give us what we want.

For example, the least fixed point of Φ_{HF} does not contain any non-well-founded sets, such as Ω. Intuitively, however, Ω should be counted as an hereditarily finite set: every set in its transitive closure is a singleton.

The set Ω *is* in the largest fixed point of Φ_{HF}. So, perhaps the largest fixed point of Φ_{HF} contains exactly those sets that we, intuitively, would call hereditarily finite. The following observation gives evidence for this:

Proposition 1.16 A set is in the largest fixed point of Φ_{HF} iff it is in the solution set of a system of equations $(x = s_x)_{x \in \mathcal{X}}$ in which each s_x is finite. □

The fact that a set s is in the solution set of a system of equations such that each s_x is finite means that each set in the transitive closure of s has finitely many elements. The elements of the largest fixed point of Φ_{HF} are exactly the sets with a transitive closure that contains finite sets only, which means that it is the largest fixed point of Φ_{HF} rather than the least fixed point that is the class of all hereditarily finite sets in the non-well-founded universe.

A lot of technical work in non-well-founded set theory is concerned with the relation between largest fixed points of continuous operators and solution sets of systems of equations of a certain form. It turns out that for operators that are 'uniform,' there is a natural relation: if Φ is a uniform operator, then the largest fixed point of Φ is precisely the solution set of a system of equations that is a Φ-coalgebra:

Definition 1.17 (Φ-coalgebra)[5]

A system of equations $(x = s_x)_{x \in \mathcal{X}}$ is a Φ-*coalgebra* just in case $s_x \in \Phi(\mathcal{X})$ for each $x \in \mathcal{X}$, and the set $\mathcal{X}$ is *new* for Φ. □

A Φ-coalgebra is a generalised system of equations of a certain form. We know that each system of equations has a unique solution. It turns out that when Φ is a uniform operator, the solution set of each Φ-coalgebra is a subset of the largest fixed point of Φ. Moreover, the largest fixed point is exactly the union of all solution sets of Φ-coalgebras.

Proposition 1.18 (representation theorem for largest fixed points)

For each uniform operator Φ, it holds that the largest fixed point of Φ is the union of all solution sets of Φ-coalgebras. □

This is a useful result, because it tells us precisely what kind of sets are in the largest fixed point of an operator. Proposition 1.16 is a special case of the representation theorem.

Consider, as an example, the largest fixed point of $\Phi_{\mathcal{L}}$. Since $\Phi_{\mathcal{L}}$ is set-continuous, we know that such a fixed point exists. Let us call it $\mathcal{L}^*$. The representation theorem tells us that all and only solutions of systems of equations $(x = s_x)_{x \in \mathcal{X}}$ such that each $s_x \in \Phi_{\mathcal{L}}(\mathcal{X})$ are sentences of $\mathcal{L}^*$. Consider the following generalised system of equations:[6]

$$\begin{aligned} x &= p \wedge q \\ y &= p \wedge y \\ z &= \neg z \end{aligned}$$

[5]This proposition contains the notions of 'uniform operation' and 'a set of indeterminates that is new for Φ.' I will not properly define these conditions here, but only mention that all the operators defined in this dissertation are uniform. All sets of indeterminates are new for Φ_{HF} and Φ_{ω}, and all sets that do not contain elements of $\mathcal{P} \cup \{\wedge, \neg\}$ are new for $\Phi_{\mathcal{L}}$. In essence, the condition that operators are uniform guarantees that there are 'enough' new indeterminates to find proper systems of equations to represent the largest fixed point. Choosing new indeterminates in the equations that represent a largest fixed point ensures that the choice of the indeterminates does not interfere with the effect of the operator itself.

[6]Remember that $s_1 \wedge s_2$ is an abbreviation for $\langle s_1, \wedge, s_2 \rangle$.

By the solution lemma, we know that this system has a unique solution, say δ. The representation theorem tells us that δ_x, δ_y and δ_z are elements of $\mathcal{L}^*$. The sentence δ_x is simply the familiar sentence $p \wedge q$. The other two sentences are more interesting. It holds that $\delta_y = \delta_x \wedge \delta_y$, i.e. δ_y is the unique "sentence" ϕ such that $\phi = p \wedge \phi$. Similarly, δ_z is the unique sentence that is equal to its negation.

We will now turn to an example of a definition of a largest fixed point that will play a central role in this dissertation: the class of possibilities.

1.2 Possibilities

In this dissertation, I will use possibilities to give a semantics for epistemic logic. They are defined as follows:

Definition 1.19 (possibilities)
Let $\mathcal{A}$, a set of agents, and $\mathcal{P}$, a set of propositional variables, be given. The class of *possibilities* is the largest class such that:

- A possibility w is a function that assigns to each propositional variable $p \in \mathcal{P}$ a truth value $w(p) \in \{0, 1\}$ and to each agent $a \in \mathcal{A}$ an information state $w(a)$.
- An information state σ is a set of possibilities. □

A possibility w specifies which propositions are true and which are false, and it characterizes the information that each of the agents has in the form of an information state σ, that consists of the set of possibilities the agent considers possible in w. A very similar model has been proposed by Aczel for modeling transition systems. Barwise and Moss do similar things when they show how one can interpret a unimodal language on hypersets *simpliciter*. In chapter 2, I will study the relations between possibilities and other models for epistemic logic in more detail. Here, I will study the set theoretical background behind this definition of possibilities, and show how one can define operations on possibilities using the set-theoretical machinery developed in Aczel (1988) and Barwise and Moss (1996).

Consider the following operation on sets:

$$\Phi_{\mathsf{Poss}}(s) = \{f \mid f \text{ is a function } \mathcal{P} \mapsto \{0,1\} \cup \mathcal{A} \mapsto \mathsf{Pow}(s)\}$$

It can easily be checked that this operation is monotone. That means that Φ_{Poss} has both a least and a largest fixed point. It is in particular the largest fixed point of Φ_{Poss} that interests us here, since this class turns out to be exactly the class of possibilities. In the following, I will write Poss for the largest fixed point of Φ_{Poss}.

To see that Poss is the class of all possibilities, consider the representation theorem for largest fixed points, proposition 1.18. This proposition tells us that the elements of the largest fixed point of Φ_{Poss} are exactly the solutions of certain systems of equations, namely all those systems $(x = s_x)_{x \in \mathcal{X}}$ such that each $s_x \in \Phi_{\mathsf{Poss}}(\mathcal{X})$.[7]

Suppose that $\mathcal{P}$ is a singleton set containing p only, and there is only one agent a in $\mathcal{A}$. If we represent functions by the set of their argument-value pairs, then the following is an example of a simple Φ_{Poss}-coalgebra:

$$x = \{(p, 0), (a, \emptyset)\}$$

The solution of this system of equations assigns to x the function that assigns to p the value 0, and to a the information state that is the empty set. Clearly, this is a possibility in the sense of definition 1.19. Note also that the solution of x is a function in the well-founded universe (in fact, it is an element of the least fixed point of Φ_{Poss}).

As a second example, consider the following system:

$$\begin{aligned} x &= \{(p, 0), (a, \{x, y\})\} \\ y &= \{(p, 1), (a, \{y\})\} \end{aligned}$$

It is not hard to see that this is a Φ_{Poss}-coalgebra, which means that the solutions for x and y are possibilities. Let δ be this solution. Then δ_x is the (unique) function w such that $w(p) = 0$, and $w(a) = \{w, v\}$, while δ_y is the function v such that $v(p) = 1$ and $v(a) = \{v\}$.

The largest fixed point Poss of Φ_{Poss} contains all and only the solutions of systems of equations that describe functions that assign to each propositional variable a truth-value and to each agent a set of such functions. So, Poss contains exactly those set-theoretical objects that are possibilities in the sense of definition 1.19.

In this dissertation, I will often define particular examples of possibilities using a format that is more natural for functions than a Φ_{Poss}-coalgebra. Consider the following equations.

$$\begin{aligned} w(p) &= 0 \\ w(a) &= \{w, v\} \\ v(p) &= 1 \\ v(a) &= \{v\} \end{aligned}$$

The following proposition shows that we can indeed use equations such as these to define possibilities.

[7] The operation Φ_{Poss} is uniform, as required. All sets of indeterminates that do not include elements of either $\mathcal{P}$ or $\mathcal{A}$ are new for Φ_{Poss}.

Definition 1.20 (equations for possibilities)

Given $\mathcal{A}$ and $\mathcal{P}$, a *system of equations for possibilities* in the indeterminates $\mathcal{X}$ is a set of equations such that for each $x \in X$ there is for each $p \in \mathcal{P}$ exactly one equation of the form $x(p) = i$, where $i \in \{0,1\}$, and for each $a \in \mathcal{A}$, the system contains exactly one equation of the form $x(a) = \sigma$, where $\sigma \subseteq X$.

A *decoration of a system of equations for possibilities* is a function δ that assigns to each indeterminate x a possibility δ_x in such a way that if $x(p) = i$ is an equation then $\delta_x(p) = i$, and if $x(a) = \sigma$ is an equation, then $\delta_x(a) = \{\delta_y \mid y \in \sigma\}$. □

Proposition 1.21 It holds that:

Poss is precisely the union of all solution sets of systems of equations for possibilities. □

We found it useful to depict sets with graphs. However, if we restrict our attention to possibilities, the kind of graphs we used to depict sets are a bit cumbersome to work with. To picture possibilities it is easier to use labeled graphs.

Definition 1.22 (labeled graph)

Given a set of propositional variables $\mathcal{P}$ and a set of agents $\mathcal{A}$, a *labeled graph* is a triple $\mathcal{G} = (G, (\xrightarrow{a})_{a\in\mathcal{A}}, V)$, where G is a set of nodes, for each $a \in \mathcal{A}$, $\xrightarrow{a}$ is a relation over G, and V is a function that assigns to each $p \in \mathcal{P}$ a subset of G.

A *pointed labeled graph* is a pair $(\mathcal{G}, n)$, where $\mathcal{G}$ is a labeled graph, and $n \in G$. □

Figure 1.5 depicts a labeled graph. A labeled graph is just a Kripke model. I will

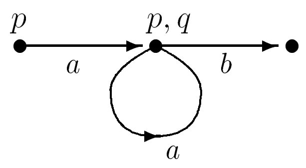


Figure 1.5.

discuss Kripke semantics in more detail in the following chapters.

The relation between labeled graphs and possibilities is very similar to that between graphs and sets: we can decorate labeled graphs with possibilities.

Definition 1.23 (decorating labeled graphs)

If $\mathcal{G}$ is a labeled graph, then a function δ that assigns to each node of $\mathcal{G}$ a possibility is a *decoration* just in case it holds for each node n that:

1. $\delta_n(p) = 1$ iff $n \in V(p)$

2. $\delta_n(a) = \{\delta_m \mid n \xrightarrow{a} m\}$ $\square$

Intuitively, a node n in a labeled graph can be decorated with a possibility w just in case w assigns the value 1 to a propositional variable p when $n \in V(p)$, and w assigns to each agent the set of nodes that can be reached from n by an arrow labeled with a.

For example, consider figure 1.5, and consider the three possibilities w, v and u specified by the following equations:

$$\begin{array}{lll} w(p) = 1 & v(p) = 1 & u(p) = 0 \\ w(q) = 0 & v(q) = 1 & u(q) = 0 \\ w(a) = \{v\} & v(a) = \{v\} & u(a) = \emptyset \\ w(b) = \emptyset & v(b) = \{u\} & u(b) = \emptyset \end{array}$$

If δ is the unique solution (given by the solution lemma) to this system of equations, then the decoration of the graph pictured that assigns w to the left-most node, v to the middle node and u to the right-most node is a decoration of the graph of figure 1.5

The results about the relation between labeled graphs and possibilities are analogous to the relation between graphs and sets:

Proposition 1.24

- Each possibility can be pictured by a labeled graph.
- Each labeled graph has a unique decoration.
- Two labeled graphs have the same decoration iff they are bisimilar. $\square$

This proposition gives us yet another characterization of what a possibility is: a possibility is anything that can be pictured by a labeled graph.

1.2.1 Operations on Possibilities

A standard tool for defining operations in well-founded set-theory is by *recursion*: given a ordered set $(A, <)$ of a set (or class) A that is well-founded[8] we can define a function f on A by recursion by "following the ordering of A"; i.e. by defining $f(s)$ in terms of the values of f on elements of $\{t \mid t < s\}$.

This technique is particularly useful for defining functions on the least fixed points of operators Φ, because they come with a natural ordering that is given by their characterization as $\bigcup_\alpha \Phi_\alpha$ (cf. proposition 1.14). For example, one can well-order the least fixed point of an operator Φ by letting $s < t$ iff for all α, if $t \in \Phi_\alpha$ then there is a $\beta < \alpha$ such that $s \in \Phi_\beta$. For the operator Φ_ω, this gives us

[8]An ordered set $(A, <)$ is well-founded iff there are no infinite sequences $s_0, s_1, s_2 \ldots$ of elements of A such that $s_0 > s_1 > \ldots$.

the usual 'smaller than' ordering on the natural numbers, and for the operator $\Phi_{\mathcal{L}}$, it holds that if ϕ is a (proper) subformula of ψ, then $\phi < \psi$.

A typical example of a definition by recursion is the following operation f on elements of the least fixed point of $\Phi_{\mathcal{L}}$ that substitutes $\perp$ for all propositional variables in a formula:

$$f(p) = \perp \text{ for each propositional variable } p$$
$$f(\neg\phi) = \neg f(\phi)$$
$$f(\phi \wedge \psi) = f(\phi) \wedge f(\psi)$$

Unfortunately, when we want to define functions on sets that are not least, but largest fixed points of operators, this technique cannot be used straightforwardly. Largest fixed points are, in general, not naturally ordered in the way that least fixed points are. So we need another technique to define operations on largest fixed points.

The corecursion theorem of Barwise and Moss (1996) gives us a general way of defining functions from an arbitrary class *into* largest fixed points. So in a sense, this seems just the opposite of what we need: by recursion we can define functions that have a least fixed point as their domain, corecursive definitions define functions that have a largest fixed point as their range. But in this dissertation we will be interested only in operations that have possibilities in both their domain and their range, and the corecursion theorem is very useful for defining such operations.

The following fact is a special case of the much more general corecursion theorem in Barwise and Moss. I have tried to formulate it in such a way that it is just general enough to cover the kind of functions that will be defined in this dissertation.

Proposition 1.25 (corecursion for possibilities)
Let Φ_{Poss} be the operator defined on page 12, whose largest fixed point is the class of all possibilities. Let C be any class, and let π be a function $\pi : C \mapsto \Phi_{\mathsf{Poss}}(C)$.

There is a unique function ϕ that maps elements of C into the largest fixed point of Φ_{Poss} such that

$$\begin{array}{rcl} \phi(c)(p) & = & \pi(c)(p) \text{ for each } p \in \mathcal{P} \\ \phi(c)(b) & = & \{\phi(c') \mid c' \in \pi(c)(b)\} \end{array}$$

If ϕ is defined corecursively in this way, we call the function π the *pump* of the definition. □

In this dissertation, I will often make implicit use of the corecursion theorem. I will give some examples here.

Consider first the simple operation on possibilities that resets the value of one distinguished propositional q to the value 1.

$\pi(w)$ is the possibility that is just like w, except that $\pi(w)(q) = 1$

Since the class of possibilities is a fixed point of Φ_{Poss}, the function π maps a class C (namely the class of possibilities) into $\Phi_{\mathsf{Poss}}(C)$. That means that we can use π as the pump of a definition of a function ϕ. This will be a function from possibilities into possibilities such that for each w, $\phi(w)$ is the possibility such that:

$$\begin{array}{rcl} \phi(w)(q) & = & 1 \\ \phi(w)(p) & = & w(p) \text{ if } p \neq q \\ \phi(w)(a) & = & \{\phi(v) \mid v \in w(a)\} \end{array}$$

The result of applying the function ϕ to a possibility w changes the value of the propositional variable q to 1 in all possibilities in the transitive closure of w. If we restate this in terms of pictures of w, it means that we get a picture of $\phi(w)$ by taking a picture of w, and changing the valuation function V in such a way that $V(q)$ is the set of all nodes of the picture.

For a more interesting example, we temporarily step out of the realm of epistemic semantics and consider the semantics of process algebra. Aczel defines structures as a model for process algebra that are similar to possibilities, as an alternative to the semantics that makes use of graphs and bisimulation. (Since process algebra is not the topic of this dissertation, these remarks have to suffice.) If we assume that our set of propositional variables is empty (which is a fairly natural assumption in process algebra), we can define the operation $||$ of 'free merge' between two possibilities by setting, for each $s \in \mathcal{A}$:

$$(w||v)(a) = \{(w'||v) \mid w' \in w(a)\} \cup \{(w||v') \mid v' \in v(a)\}$$

We can show that this is well-defined by giving a corecursive definition of a function that satisfies the equation above. The pump of such a corecursive definition is the function $\pi : \mathsf{Poss} \times \mathsf{Poss} \mapsto \Phi_{\mathsf{Poss}}(\mathsf{Poss} \times \mathsf{Poss})$ given by:

$$\pi(\langle w, v\rangle)(a) = \{\langle w', v\rangle) \mid w' \in w(a)\} \cup \{\langle w, v'\rangle \mid v' \in v(a)\}$$

The following construction will play an important role in this dissertation. Given a relation R on possibilities, and given a subset $\mathcal{B}$ of $\mathcal{A}$, we define a new function $\phi_{\mathcal{B}}$ on possibilities such that:

$$\begin{array}{rcl} \phi_{\mathcal{B}}(w)(p) & = & w(p) \text{ for each } p \in \mathcal{P} \\ \phi_{\mathcal{B}}(w)(a) & = & \left\{ \begin{array}{ll} \{\phi_{\mathcal{B}}(v) \mid \exists w' \in w(a) \text{ such that } w'Rv\} & \text{if } a \in \mathcal{B} \\ w(a) & \text{if } a \notin \mathcal{B} \end{array} \right. \end{array}$$

If we apply the function $\phi_{\mathcal{B}}$ to a possibility w, the result is a possibility v that is just the same as w in the values it assigns to propositional variables and in the information states that it assigns to agents not in $\mathcal{B}$. For agents in $\mathcal{B}$, the new information state $v(a)$ of a is obtained by collecting all the possibilities that are connected by R to some possibility in $w(a)$, and then applying the function ϕ to these possibilities. I will use such functions to model the effect of a 'conscious update' in chapter 4.

Using the corecursion theorem, we can show that this is well-defined. Given a relation R, we define a pump $\pi : \mathsf{Pow}(\mathcal{A}) \times \mathsf{Poss} \mapsto \Phi(\mathsf{Pow}(\mathcal{A}) \times \mathsf{Poss})$ by:

$$\begin{array}{rcl} \pi(\langle \mathcal{B}, w\rangle)(p) & = & w(p) \\ \pi(\langle \mathcal{B}, w\rangle)(a) & = & \left\{ \begin{array}{ll} \{\langle \mathcal{B}, v'\rangle \mid \exists w' \in w(a) \text{ such that } w'Rv'\} & \text{if } a \in \mathcal{B} \\ \{\langle \emptyset, w'\rangle \mid w' \in w(a)\} & \text{if } a \notin \mathcal{B} \end{array} \right. \end{array}$$

Using the function π as a pump, we get a function $\phi : \mathsf{Pow}(\mathcal{A}) \times \mathsf{Poss} \mapsto \mathsf{Poss}$. We can now define the function $\phi_{\mathcal{B}}$ by

$$\phi_{\mathcal{B}}(w) = \phi(\mathcal{B}, w)$$

To see that this is indeed the function $\phi_{\mathcal{B}}$ that we are looking for, note first that that $\phi_{\mathcal{B}}(w)(p) = \pi(\mathcal{B}, w)(p) = w(p)$. If $a \in \mathcal{B}$, then $\phi_{\mathcal{B}}(w)(a) = \{\phi(\mathcal{C}, v') \mid (\mathcal{C}, v') \in \pi(\mathcal{B}, w)(a)\} = \{\phi(\mathcal{B}, v') \mid \exists w' \in w(a) : w'Rv'\} = \{\phi_{\mathcal{B}}(v') \mid \exists w' \in w(a)$ such that $w'Rv'\}$. For the case in which $a \notin \mathcal{B}$, observe first that $\phi(\emptyset, w) = w$ (this can be shown by showing that the two are bisimilar) for each w. Since $\phi_{\mathcal{B}}(w)(a) = \{\phi(\emptyset, w') \mid w' \in w(a)\}$, it holds that $\phi_{\mathcal{B}}(w)(a) = w(a)$.

This concludes my discussion of non-well-founded set-theory. In the next chapter, we will study the relation of bisimulation and its relation to models for epistemic logic in more detail. In the chapter after that, we turn to the study of epistemic logic proper, and use possibilities as a semantics for epistemic logic.

2

Bisimulation and Bounded Bisimulation

Ever since Kripke (1963b), labeled graphs are the classical way of providing modal logic with a model. In this dissertation, and in Barwise and Moss (1996), non-well-founded sets are used as a semantics for modal logic. We have seen in the previous chapter how the relation of bisimulation provides a connection between these two kinds of structures.

In this chapter, such interrelationships are further explored with the help of the concept of bounded bisimulation. I will define a bounded bisimulation relation for each ordinal, and show that the sequence of bounded bisimulations that is defined in this way converges to bisimulation. It turns out that bounded bisimulations provide us with a natural way to explicate the relation between Kripke models and so-called 'knowledge structures' that were introduced in the 1980s as an alternative to Kripke Semantics. Furthermore, I will show how bisimulation corresponds with equivalence in infinitary modal logic, and how bounded bisimulation corresponds to fragments of infinitary modal logic that consist of sentences up to a certain modal depth.

The chapter is organized as follows. The next section contains definitions of bisimulation and bounded bisimulation on Kripke models, which are the core notions in this chapter. The next two sections contain definitions of non-well-founded models for modal logic, and of knowledge structures. It is shown how these models can be seen as representing bisimulation classes and bounded bisimulation classes of Kripke models respectively. In the two sections after that we show how bounded bisimulation can be characterized by certain fragments of infinitary modal logic and by Ehrenfeucht games. The last section contains some examples of results proven by different authors in different contexts and shows how they can be seen as closely related using the notions introduced in this chapter.

This chapter is based on Gerbrandy (1997a). Not all of the definitions and theorems are originally mine, and I have tried to give credit as much as possible.

2.1 Bisimulation and Bounded Bisimulation

I start with repeating two definitions from the previous chapter: the definition of a Kripke model, and that of bisimulation between models.

A Kripke model is the same sort of object as the labeled graph of the previous section.

Definition 2.1 (Kripke models)
Let a set of propositional variables $\mathcal{P}$ be given. A *Kripke model* K is a triple $(W, \longrightarrow, V)$, where W is a set of indices, $\longrightarrow$ is a relation on W, and $V : \mathcal{P} \mapsto \text{Pow}(W)$ is a valuation function that assigns to each atomic sentence in $\mathcal{P}$ a subset of W.
A *pointed Kripke model* is a pair (K, w), where K is a Kripke model $(W, \longrightarrow, V)$, and $w \in W$. □

Where it does not lead to confusion, I will often write w instead of (K, w). I have defined here a Kripke model with a single accessibility relation. When Kripke models are used as models for multi-agent epistemic logic, it is often useful to have more than just a single accessibility relation in the model. Most of the definitions and results in this chapter can easily be generalized to apply to models with an arbitrary (finite) number of accessibility relations.

Bisimulation was first defined in the context of modal logic by Van Benthem (1976), and re-invented later in the context of transition systems by Park (1981).

Definition 2.2 (bisimulation)
A relation $R \subseteq W \times W'$ is a *bisimulation* between two models $K = (W, \longrightarrow, V)$ and $K' = (W', \longrightarrow', V')$ iff for all $w \in W$ and $w' \in W'$, if wRw' then:

1. $w \in V(p)$ iff $w' \in V'(p)$ for all $p \in \mathcal{P}$.
2. For all v such that $w \longrightarrow v$, there is a v' such that $w' \longrightarrow' v'$ and vRv'.
3. For all v' such that $w' \longrightarrow' v'$, there is a v such that $w \longrightarrow v$ and vRv'.

Two Kripke models (K, w) and (K', w')[1] are *bisimilar*, $(K, w) \simeq (K', w')$, iff there is a bisimulation between K and K' connecting w with w'. □

All this is familiar from the previous section. Consider now the following construction:

[1] From now on, K and K' will be silently understood to be the models $(W, \longrightarrow, V)$ and $(W', \longrightarrow', V')$ respectively.

Definition 2.3 (bounded bisimulation)
Let α be an ordinal. We define the notion of α-bisimulation by induction on α: two pointed Kripke models (K, w) and (K', w') are α-bisimilar, $(K, w) \simeq_\alpha (K', w')$, iff

1. $w \in V(p)$ iff $w' \in V'(p)$ for all $p \in \mathcal{P}$.

2. For all $\beta < \alpha$, and for all v such that $w \longrightarrow v$, there is a v' such that $w' \longrightarrow' v'$ and $(K, v) \simeq_\beta (K, v')$,

3. For all $\beta < \alpha$, and for all v' such that $w' \longrightarrow' v'$, there is a v such that $w \longrightarrow v$ and $(K, v) \simeq_\beta (K', w')$. $\Box$

Two pointed models (K, w) and (K', w') are 0-bisimilar iff all propositional variables have the same truth-value w and w'. They are $\alpha + 1$-bisimilar iff the propositional variables have the same truth-values, and for each successor of w there is an α-bisimilar successor of w', and vice versa. If α is a limit ordinal, then (K, w) and (K', w') are α-bisimilar just in case they are β-bisimilar for all $\beta < \alpha$. Where it is not likely to lead to confusion, I will write $w \simeq_\alpha w'$ instead of $(K, w) \simeq_\alpha (K', w')$.

Bounded bisimulation for ordinals smaller or equal to ω have been defined in Hennessy and Milner (1985) and several other places. The generalization of the notion to arbitrary ordinals is new, as far as I know, although the work in Doets (1987) about infinitary games in first-order logic is closely related.

Bisimulation and bounded bisimulation are closely related. The following proposition shows that α-bisimulations are proper approximations of bisimulation.

Proposition 2.4 For all possibilities w and w' and ordinals α and β, it holds that:

- $w \simeq w'$ iff for all ordinals α: $w \simeq_\alpha w'$.

- If $\beta \leq \alpha$, and R is a α-bisimulation between w and w', then R is an β-simulation between w and w'.

- For each α, there are w and w' such that $w \simeq_\alpha w'$ but $w \not\simeq w'$.

proof: For the first statement, it is easily seen that any relation that is an α-bisimulation for each α is a bisimulation. For the other direction, show by induction on α that each bisimulation is an α-bisimulation.

The proof of the second statement is completely straightforward, and we will prove the third statement together with proposition 2.24. $\Box$

2.2 Languages

The whole point of defining Kripke models is that they can be used for the interpretation of a modal language. The intuitions and motivations behind the following definitions are the topic of the next chapter; here, I will concentrate mostly on certain formal aspects.

Definition 2.5 (language of infinitary modal logic)
Let a set of propositional variables $\mathcal{P}$ be given. The language of infinitary modal logic $\mathcal{L}_\infty$ is the smallest class such that: $\mathcal{P} \subseteq \mathcal{L}_\infty$, if ϕ in $\mathcal{L}_\infty$, then $\neg\phi$ and $\Diamond\phi$ are in $\mathcal{L}_\infty$, and if Φ is a subset of $\mathcal{L}_\infty$, then $\bigwedge \Phi \in \mathcal{L}_\infty$. □

The definitions of satisfaction in a model are standard.

Definition 2.6 (satisfaction of sentences of $\mathcal{L}_\infty$)

$$\begin{array}{rcl}
(K, w) \models p & \text{iff} & w \in V(p) \\
(K, w) \models \neg\phi & \text{iff} & (K, w) \not\models \phi \\
(K, w) \models \bigwedge\Phi & \text{iff} & (K, w) \models \phi \text{ for all } \phi \in \Phi \\
(K, w) \models \Diamond\phi & \text{iff} & \text{there is a } v \text{ such that } w \longrightarrow v \text{ and } (K, v) \models \phi
\end{array} \qquad \Box$$

The truth of sentences of infinitary modal logic is closely connected with the notions of bisimulation and bounded bisimulation. Two models are bisimilar exactly when they satisfy the same sentences of infinitary modal logic. Bounded bisimulation can be characterized by certain fragments of $\mathcal{L}_\infty$ in a similar way: two Kripke models are α-bisimilar just in case they satisfy the same sentences of $\mathcal{L}_\infty$ up to depth α.

Definition 2.7 (depth of formula's) Define the depth of a formula of $\mathcal{L}_\infty$ as a function from formula's to ordinals:

$$\begin{array}{rcl}
\mathsf{depth}(p) & = & 0 \\
\mathsf{depth}(\neg\phi) & = & \mathsf{depth}(\phi) \\
\mathsf{depth}(\Diamond\phi) & = & \mathsf{depth}(\phi) + 1 \\
\mathsf{depth}(\bigwedge \Phi) & = & \bigcup\{\mathsf{depth}(\phi) \mid \phi \in \Phi\}
\end{array}$$

For each ordinal α, the language $\mathcal{L}_\alpha$ is the class of all sentences of $\mathcal{L}_\infty$ of depth $\leq \alpha$. We write that $(K, w) \equiv_\alpha (K', w')$ iff (K, w) and (K', w') satisfy the same sentences of $\mathcal{L}_\alpha$. □

The following proposition states that satisfaction of sentences of $\mathcal{L}_\alpha$ corresponds exactly to α-bisimulation.

Proposition 2.8 It holds that:

$$(K, w) \simeq_\alpha (K', w') \text{ iff } (K, w) \equiv_\alpha (K', w')$$

proof:

[$\Rightarrow$] We prove by induction on the structure of sentences ϕ of $\mathcal{L}_\infty$ that if ϕ is of depth α, $w \models \phi$ and $w \simeq_\alpha w'$, then $w' \models \phi$.

The only nontrivial case is $\Diamond\phi$. Note that the depth of a formula of the form $\Diamond\phi$ is always a successor ordinal. So, let $\mathsf{depth}(\Diamond\phi) = \alpha + 1$, and assume that $w \simeq_{\alpha+1} w'$. If $w \models \Diamond\phi$ there must be a v such that $w \longrightarrow v$ and $v \models \phi$. Since w and w' are $\alpha + 1$-bisimilar, there is a v' such that $w' \simeq_\alpha v'$ and $w' \longrightarrow' v'$. Because $\mathsf{depth}(\Diamond\phi) = \alpha + 1$, it must hold that $\mathsf{depth}(\phi) = \alpha$, and we can use the induction hypothesis to conclude that $w' \models \Diamond\phi$.

[$\Leftarrow$] We show by induction on α that $\equiv_\alpha$ is an α-bisimulation.

$\alpha = 0$: immediate.

$\alpha + 1$: Assume $w \equiv_{\alpha+1} w'$. Take any v such that $w \longrightarrow v$. Assume to the contrary that for no v' such that $w' \longrightarrow v'$ it holds that $v \simeq_\alpha v'$. Then, by induction hypothesis, for each v' such that $w' \longrightarrow v'$ there is a $\phi_{v'} \in \mathcal{L}_\alpha$ such that $v \models \phi_{v'}$ but $v' \not\models \phi_{v'}$. Then $w \models \Diamond \bigwedge\{\phi_{v'} \mid w' \longrightarrow v'\}$, which is an $\mathcal{L}_{\alpha+1}$-sentence. But $w' \not\models \Diamond \bigwedge\{\phi_{v'} \mid w' \longrightarrow v'\}$, contradicting our assumption.

The final case in the induction is when $w \equiv_\lambda w'$ for some limit ordinal λ. In that case, $w \equiv_\beta w'$ for all $\beta < \lambda$. So, by induction hypothesis, $w \simeq_\beta w'$ for all $\beta < \lambda$, from which it follows by definition of λ-bisimulation that $w \simeq_\lambda w'$. □

The language $\mathcal{L}_\omega$ is the classical fragment of $\mathcal{L}_\infty$. It is a well-known result that the language $\mathcal{L}_\omega$ cannot distinguish between two ω-bisimilar models. Also, the fact that $\mathcal{L}_\infty$ corresponds to bisimulation is well-known. We get these results as corollaries of the above:

Corollary 2.9 (finitary and infinitary languages)[2]

- $(K, w) \simeq_\omega (K', w')$ iff for all $\phi \in \mathcal{L}_\omega$: $(K, w) \models \phi$ iff $(K', w') \models \phi$
- $(K, w) \simeq (K', w')$ iff for all $\phi \in \mathcal{L}_\infty$: $(K, w) \models \phi$ iff $(K', w') \models \phi$ □

What these results tell us is that as far as modal logic is concerned, the distinctions between bisimilar models are inessential. In the next sections, we will look at two other ways of giving a semantics to modal logic where the differences between bisimilar models are conflated.

2.3 Sets

There have been proposals to 'simplify' Kripke semantics by identifying bisimilar Kripke models. Examples are the model for process algebra of Aczel (1988) and the possibilities that are used as models for epistemic logic in this dissertation. This idea finds its most elegant expression in the use of sets *simpliciter* as a model for modal logic in Barwise and Moss (1996).

[2]These results seem to be 'folklore.' For a reference, take Van Benthem (1991)

Consider the class of non-well-founded sets based on a set $\mathcal{P}$ of urelements.[3] We can interpret a modal language that has $\mathcal{P}$ as it atomic variables using sets as models. If an atomic sentence p is an element of a set, it is true in that set, otherwise, it is false. The modal operator is interpreted relative to the 'element-of' relation between sets.

Definition 2.10 Let a be a set (possibly containing urelements from $\mathcal{P}$).

$$\begin{array}{rcl} a \models p & \text{iff} & p \in a \\ a \models \neg\phi & \text{iff} & a \not\models \phi \\ a \models \bigwedge \Phi & \text{iff} & a \models \phi \text{ for each } \phi \in \Phi \\ a \models \Diamond\phi & \text{iff} & \text{there is a } b \in a \text{ such that } b \models \phi \end{array} \qquad \Box$$

Just as we decorated multi-modal Kripke models with possibilities in the previous chapter, we can decorate Kripke models with sets.

Definition 2.11 Let (K, w) be a pointed Kripke model.

A *decoration* δ of K is a function δ that assigns to each index w in K a set $\delta_w = \{p \mid w \in V(p)\} \cup \{\delta_v \mid w \longrightarrow v\}$.

If δ is a decoration of K and w is in K, then δ_w is a *solution* of (K, w), and (K, w) is a *picture* of δ_w. $\Box$

The following holds:

Proposition 2.12

- Each pointed Kripke model has exactly one set as its solution.
- Each set has a model as its picture.
- Two models are bisimilar iff they are pictures of the same set.
- If (K, w) is a picture of a, then for each $\phi \in \mathcal{L}_\infty$: $(K, w) \models \phi$ iff $a \models \phi$.

proof: A variation of the first three items was proven in proposition 1.6. The last item can be proven by a straightforward induction on the structure of ϕ. $\Box$

Since sets and their pictures satisfy the same sentences of modal logic, this means that we can use sets instead of Kripke models as a semantics of modal logic. In a way, this provides us with a simplification of the semantics of modal logic: instead of a whole class of bisimilar models that all satisfy the same sentences anyway, we can use the set that is the solution of each of these models to represent the bisimulation class.

[3]These are all the sets that can be proven to exist in ZFC(AFA) together with an axiom that says that for any element p of $\mathcal{P}$, there is a set containing p.

2.4 Knowledge Structures

Knowledge structures are models for epistemic logic that were introduced in Fagin and Vardi (1985) and in Fagin et al. (1991) as a computationally more attractive alternative to Kripke models. The relationship between the knowledge structures and Kripke models has been studied by Hamilton and Delgrande (1989) and in more detail by Fagin (1994). The main result of this section is that knowledge structures can be identified with α-bisimulation classes of Kripke models in the same way as sets can be seen as representing bisimulation classes.

Definition 2.13 (knowledge structures)
For each ordinal α, an *α-knowledge structure* is a function f on $\mathcal{P} \cup \alpha$ such that $f(p) \in \{0, 1\}$ for each $p \in \mathcal{P}$, and for each $\beta < \alpha$, $f(\beta)$ is a set of β-knowledge structures. We let $h_{<\beta}$ be the restriction of h to $\mathcal{P} \cup \beta$.

Each α-knowledge structure f has to satisfy *extension*, which is the following requirement:

- For all ordinals β and γ such that $\beta < \gamma < \alpha$ it holds that $g \in f(\beta)$ iff there is an $h \in f(\gamma)$ such that $g = h_{<\beta}$. □

The intuition behind this definition is that an α-knowledge structure represents the information of an agent “up to depth α.” A 0-knowledge structure simply assigns a truth-value to each of the propositional variables. A 1-knowledge structure does the same, but also assigns to the ordinal 0 a set of 0-knowledge structures. This set represents the information of an agent up to depth 0. A 2-knowledge structure assigns to the number 1 a set of 1-knowledge structures that represents the information of that agent up to depth 1, and to the number 0 a set of knowledge structures that represents the information up to depth 0. The requirement that knowledge structures satisfy extension guarantees that the knowledge as represented up to depth 0 is compatible with the knowledge up to depth 1.

Fagin et al. (1991) use a slightly more complex notion of a knowledge structure: they consider models for more more than one agent, and apart from extension, they postulate two other restrictions. These complications are irrelevant for the purposes of this section, however.

Satisfaction of a sentence ϕ in am α-knowledge structure is only defined if the depth of ϕ is lower or equal to α. It can be defined as follows:

Definition 2.14 (satisfaction of sentences in knowledge structures)
The satisfaction relation $\models$ is a relation between α-knowledge structures f and sentences ϕ such that $\mathsf{depth}(\phi) \leq \alpha$, defined by induction on ϕ as follows:

$$\begin{array}{rll} f \models p & \text{iff} & f(p) = 1 \\ f \models \neg\phi & \text{iff} & f \not\models \phi \\ f \models \bigwedge \Phi & \text{iff} & f \models \phi \text{ for each } \phi \in \Phi \\ f \models \Diamond\phi & \text{iff} & \text{there is a } g \in f(\mathsf{depth}(\phi)) \text{ such that } g \models \phi \end{array}$$

The satisfaction relation suggests the following relation of correspondence between knowledge structures and Kripke models.

Definition 2.15 Let (K, w) be a model. By induction, we specify for each ordinal $\alpha \geq 0$ which α-knowledge structure f corresponds with (K, w).

- The 0-knowledge structure corresponding with (K, w) is the function f on $\mathcal{P}$ such that $f(p) = 1$ iff $w \in V(p)$, and $f(p) = 0$ otherwise.

- Let g be the α-knowledge structure corresponding with (K, w). The $\alpha + 1$-knowledge structure f that corresponds with (K, w) is such that $f_{<\alpha} = g$ and $f(\alpha)$ is the set of α-knowledge structures that correspond with (K, v) for each v such that $w \longrightarrow v$.

- If λ is a limit ordinal, then the λ-knowledge structure f that corresponds with (K, w) is such that for each $\beta < \lambda$: $f_{<\beta}$ is the β-knowledge structure corresponding with (K, w). $\square$

It is not hard to check that all this is well-defined, and extension is satisfied where it needs to be. Basically the same construction can be found in Fagin (1994).

We now have the following relation between knowledge structures and Kripke models, and we have a new characterization for α-bisimulation.

Proposition 2.16 For each ordinal α:

- Each Kripke model corresponds with a unique α-knowledge structure.

- Each α-knowledge structure corresponds with a Kripke model.

- $(K, w) \simeq_\alpha (K', w')$ iff (K, w) and (K', w') correspond with the same α-knowledge structure.

- For each α, each $\phi \in \mathcal{L}_\alpha$ (cf. definition 2.7) and each Kripke model (K, w) it holds that if f is the α-knowledge structure that corresponds with (K, w), then $(K, w) \models \phi$ iff $f \models \phi$.

proof: All these results are proven by Fagin (1994), except for the claim about α-bisimulation, which has a straightforward proof. $\square$

This shows that knowledge structures and bounded bisimulation equivalence classes are related in the same way as non-well-founded models and bisimulation equivalence classes are.

2.5 Ehrenfeucht Games

In this section, I will give a game-theoretic interpretation of bounded bisimulation. The relation between game theory and bisimulation has been often noted, and studied well, for example in Stirling (1996), in Nielsen and Clausen (1994) and in Rosen (1995). Perhaps most closely related is the work of Doets (1987), who defines similar games in his work on partial isomorphism in first-order logic.

Definition 2.17 We define an *Ehrenfeucht game bounded by* α (or the α-game for short) for each ordinal α by induction on α. An α-game is played with two pointed models (K, w) and (K', w'), by two players (they are called the *spoiler* and the *duplicator*).

A 0-game simply consists of the spoiler choosing w and the duplicator answering w'. The duplicator wins iff for each p, $w \in V(p)$ iff $w' \in V'(p)$.

An $\alpha + 1$-game is played as follows: the spoiler must choose a u from one of the models such that either $w \longrightarrow u$ or $w' \longrightarrow u$, and the duplicator must reply with an index u' from the other model. The players then proceed to play the α-game on (K, u) and (K', u') (or on (K, u') and (K', u), depending on the model that the spoiler chose). The duplicator wins if (1) $V(w) = V'(w')$, (2) he chose u' such that $v \longrightarrow u'$ and (3) he wins the α-game on (K, u) and (K', u').

For limit ordinals α, the spoiler starts by choosing any ordinal $\beta < \alpha$, and the players play the β-game. The duplicator wins iff he wins this β-game. □

Note that, interestingly, each of these bounded games will end after a finite number of moves. The following proposition shows that α-games can be used to characterize α-bisimulation:

Proposition 2.18 For each ordinal α:
The duplicator has a winning strategy for the α-game on (K, w) and (K', w') iff $(K, w) \simeq_\alpha (K', v)$.

proof:

[$\Rightarrow$] We prove this by induction on α. If $\alpha = 0$ we have immediately that $(K, w) \simeq_0 (K', w')$. Assume that duplicator has a winning strategy for the $\alpha + 1$-game. Then, for each successor v of w there must be a successor v' of w' such that the duplicator has a winning strategy for the α-game on (K, v) and (K', v'). Mutatis mutandis, the same holds for each successor of w'. So, by induction hypothesis and the definition of $\alpha+1$-bisimulation, (K, w) and (K', w') are $\alpha+1$-bisimilar. If α is a limit ordinal, then the duplicator has a winning strategy for the α game just in case for each $\beta < \alpha$, he has a winning strategy for each β-game. But then, by induction hypothesis, $(K, w) \simeq_\beta (K', w')$ for each $\beta < \alpha$, and hence, $(K, w) \simeq_\alpha (K', w')$.

[$\Leftarrow$] We show by induction on α that if $(K, w) \simeq_\alpha (K', w')$, then the duplicator has a winning strategy for the α-game on (K, w) and (K', w'). If $\alpha = 0$, the result is immediate. For an $\alpha+1$-game, assume that $(K,w) \simeq_{\alpha+1} (K',w')$, and the spoiler chooses v from K in his second move. Then $w \longrightarrow v$ by the rules of the game, so, since $(K,w) \simeq_{\alpha+1} (K',w')$, the duplicator can find a v' such that $w' \longrightarrow v'$ and (K,v) and (K',v') are α-bisimilar. By induction hypothesis, then, the duplicator has a winning strategy for the α-game played on (K,v) and (K',v'), and we are finished.

If α is a limit ordinal and $(K,w) \simeq_\alpha (K',w')$ then $(K,w) \simeq_\beta (K',w')$ for each $\beta < \alpha$. By induction hypothesis, the duplicator has a winning strategy for each β-game, for $\beta < \alpha$. So no matter which β is chosen by the spoiler, the duplicator wins the β-game, and hence, he wins the α-game. □

Defining a game that corresponds with bisimulation is easy, given the above.

Definition 2.19 An unbounded game is defined simply as follows: the spoiler chooses any ordinal α, and the players play the α-game. □

Proposition 2.20 The duplicator has a winning strategy for the unbounded game on (K,w) and (K',w') iff $(K,w) \simeq (K',w')$.

proof: Immediate from propositions 2.18 and 2.4. □

2.6 Some Results

This section contains some results that can be seen to be related by the interconnections noted above. One of the contributions of this section is the observation that a series of results proven by different authors using their own terminology are really manifestations of one and the same subject.

The following observation regarding the 'rank' of a model have been proven in Barwise and Moss (1996) (they speak about the 'degree' of a set) and in Fagin (1994) (who writes about the 'uniqueness ordinal' of a Kripke model).

Definition 2.21 (rank)
The *rank* of a model (K,w), $\mathsf{rank}(K,w)$, is the smallest ordinal α such that for all (K',w'): $(K,w) \simeq_\alpha (K',w')$ iff $(K,w) \simeq (K',w')$. □

The *uniqueness ordinal* of a model is defined by Fagin (1994) as the least ordinal α for which it holds that if f is the α-structure corresponding to (K,w), then for each $\beta \geq \alpha$, there is a unique β-knowledge structure g such that $g_{<\alpha} = f$. With the results of proposition 2.16, it is easily seen that the uniqueness ordinal of a model is exactly the rank of that model. Barwise and Moss (1996) define the *degree* of a set a as the least ordinal α such that $\mathcal{L}_\alpha$ characterizes a up to uniqueness. From propositions 2.12 and 2.8 it follows that the rank of a model is exactly the degree of the corresponding set. In both of the just quoted papers a result like the following is proven.

Proposition 2.22 Each model has a rank.

proof: Take any model (K, w). Since W is a set, we can use proposition 2.4 to conclude that there must be some ordinal γ such that for each w and v in K: $w \simeq_\gamma v$ iff $w \simeq v$. Let α be the smallest limit ordinal $> \gamma$. We show that for each (K', w') it holds that $(K, w) \simeq_\alpha (K', w')$ iff $(K, w) \simeq (K', w')$.

Assume $(K, w) \simeq_\alpha (K', w')$. We show that $\simeq_\alpha$ is a bisimulation between (K, w) and (K', w'). Take any v such that $w \longrightarrow v$. Since $(K, w) \simeq_\alpha (K', w')$ there must be a v' such that $v \simeq_\gamma v'$ and $w' \longrightarrow v'$. We show that $v \simeq_\alpha v'$. Since α is a limit ordinal, it is enough to show that $v \simeq_\beta v'$ for each $\beta < \alpha$.

Clearly, $v \simeq_\beta v'$ for all $\beta \leq \gamma$. So take any β such that $\gamma < \beta < \alpha$. Then, since $w \simeq_\alpha w'$, there is a u such that $u \simeq_\beta v'$. But then, $u \simeq_\gamma v$ (since $\gamma < \beta$), so $u \simeq v$ (that's how we chose γ), so $v \simeq_\beta v'$.

For the other bisimulation clause, take any v' such that $w' \longrightarrow' v'$. Then, since $\gamma < \alpha$, there is a v such that $w \longrightarrow v$ and $v \simeq_\gamma v'$. We show that $v \simeq_\alpha v'$. For take any β such that $\gamma < \beta < \alpha$. Then there is a $u \simeq_\beta v'$ such that $w \longrightarrow u$. But then $u \simeq_\gamma v$, so $u \simeq v$, and hence $v \simeq_\beta v'$. □

Corollary 2.23 Each (K, w) can be characterized up to bisimulation by a single $\mathcal{L}_\infty$-sentence (and each set can be characterized up to uniqueness by a single sentence (Barwise and Moss), and each Kripke model can be completely characterized by a single knowledge structure (Fagin)).

proof: Because the class of all α-bisimulation classes is a set, we can use proposition 2.8 and conclude that $\mathcal{L}_\alpha$ is a set, up to semantical equivalence. Let $\mathsf{rank}(K, w) = \alpha$. Then, using proposition 2.8, (K, w) is characterized up to bisimulation by its $\mathcal{L}_\alpha$-theory $\{\phi \in \mathcal{L}_\alpha \mid (K, w) \models \phi\}$. Since $\mathcal{L}_\alpha$ is, up to equivalence, a set, $\bigwedge\{\phi \in \mathcal{L}_\alpha \mid (K, w) \models \phi\}$ is a sentence characterizing (K, w) up to bisimulation. □

Proposition 2.24
For each successor ordinal α, there is a model with rank α.

proof: We will prove this proposition, and proposition 2.4 in one go. The question whether there is a model of rank α for each limit ordinal α is still open.

We will use sets instead of Kripke models to prove the proposition: in fact, we will show that each ordinal α has degree $\alpha + 1$. Using proposition 2.12, we can find a Kripke model that corresponds to α, and using the same proposition, we know that also this Kripke model has rank $\alpha + 1$.

We will show that (1) if $\alpha \leq \beta$ then $\alpha \simeq_\alpha \beta$, and that (2) if $\alpha > \beta$ then $\alpha \not\simeq_\alpha \beta$. This proves the third part of proposition 2.4. Moreover, we will show that (3) $\mathsf{rank}(\alpha) = \alpha + 1$.

(1) If $\alpha \leq \beta$, then $\alpha \simeq_\alpha \beta$.
Take any α and β such that $\alpha \leq \beta$. Then $\alpha \subseteq \beta$, so one direction of the

bisimulation clause is immediate. For the other direction, we need to show that for each $\gamma < \alpha$ and each element b of β, we can find a element a of α such that $a \simeq_\gamma b$. So take any $\gamma < \alpha$ and any $b \in \beta$. Then $b = \delta$ for some $\delta < \beta$. Distinguish two cases: (i) $\delta < \gamma$, but then δ is an element of α, and we are finished, and (ii) $\delta \geq \gamma$, but then we can simply choose γ as our witness, which is an element of α and which by induction hypothesis is γ-bisimilar to δ.

(2) If $\beta < \alpha$, then $\beta \not\simeq_\alpha \alpha$.
Take any β and α such that $\beta < \alpha$, and assume to the contrary that $\beta \simeq_\alpha \alpha$. Because $\beta < \alpha$, $\beta \in \alpha$, and so there must be an $\gamma \in \beta$ such that $\gamma \simeq_\beta \beta$. But $\gamma \in \beta$ implies that $\gamma < \beta$, while the induction hypothesis says that $\gamma \simeq_\beta \beta$ implies that $\beta \leq \gamma$.

(3) $\mathsf{rank}(\alpha) = \alpha + 1$.
Let α be any ordinal, and take any set a and assume that $a \simeq_{\alpha+1} \alpha$. Then for each element of a there is an α-bisimilar element of α and vice versa. But by induction hypothesis each $\beta \in \alpha$ is of rank smaller or equal to α, so it follows that each element of a must be bisimilar to some successor of α and vice versa. Hence, α and a are bisimilar. So, $\mathsf{rank}(\alpha) \leq \alpha + 1$.

From (1), it follows that $\mathsf{rank}(\alpha) > \alpha$, so we have shown that the rank of α is $\alpha + 1$. $\Box$

From proposition 2.4, we have the corresponding negative result that no $\mathcal{L}_\alpha$-language characterizes bisimulation:

Corollary 2.25 For each ordinal α, there are (K, w) and (K', w') such that $(K, w) \equiv_\alpha (K', w')$, but $(K, w) \not\simeq (K, v)$ (and similarly for sets) $\Box$

We now proceed to look at another result about the rank of finitely branching models.

Proposition 2.26
$\mathsf{rank}(K, w) \leq \omega$ iff (K, w) is bisimilar to a finitely branching model.

Similar results have been proven, independently as far as I know, by Alexandru Baltag in the context of non-well-founded set theory (reported in Barwise and Moss, 1996), by Fagin (1994) where the proposition is about knowledge structures, and in Hollenberg (1995), where the proposition appears as a statement about Hennessy-Milner classes.

I will first give a direct proof of proposition 2.26 (loosely based on that of Fagin, 1994), then mention the different formulations of this theorem and explain how they are related to the present one using the techniques defined in this chapter.

proof:

[$\Rightarrow$] Assume $\mathsf{rank}(K, w) \leq \omega$, but (K, w) is not bisimilar to a finitely branching model, and assume without loss of generality that w has infinitely many successors. Let f be the $\omega+1$-knowledge structure corresponding to (K, w). Since w has infinitely many non-bisimilar successors, and $\mathsf{rank}(K, w) \leq \omega$, w has infinitely many non-ω-bisimilar successors. Therefore, $f(\omega)$ contains infinitely ω-knowledge structures g (cf. proposition 2.16). Consider the function h on ω given by: $h(n) = \{g_{<n} \mid \text{there are infinitely many } g' \in f(\omega) \text{ such that } g'_{<n} = g\}$. Because there are only finitely many different n-knowledge structures for each $n < \omega$, and $f(\omega)$ is infinite, h satisfies extension, and hence it is an ω-knowledge structure.

Consider now the $\omega+1$-knowledge structures f^1 and f^2 given by $f^1_{<\omega} = f^2_{<\omega} = f_{<\omega}$, and $f^1(\omega) = f(\omega) \cup \{h\}$, $f^2(\omega) = f(\omega)/\{h\}$. Then either f^1 or f^2 is our original f, so one of them is the $\omega+1$-structure corresponding to (K, w); say it is f^1. By proposition 2.16, f^2 corresponds to a model (K', w'). Since $f^2_{<\omega} = f^1_{<\omega}$, (K, w) and (K', w') correspond to the same ω-knowledge structure, so they are ω-bisimilar. But they are not $\omega+1$-bisimilar, contradicting our assumption that $\mathsf{rank}(K, w) \leq \omega$.

[$\Leftarrow$] Assume that (K, w) is bisimilar to a finitely branching model. Take this finitely branching model, and proceed as in the proof of proposition 2.22. $\square$

Barwise and Moss formulate the following result, which they call Baltag's second theorem:

- A set a is uniquely characterized by a theory of $\mathcal{L}_\omega$ iff $a \in HF^1[\mathcal{P}]$.

The class $HF^1[\mathcal{P}]$ is the largest class of hereditary finite sets with $\mathcal{P}$ as urelements, i.e. it is the largest class of (non-well-founded) sets built from $\mathcal{P}$ such that $a \in HF^1[\mathcal{P}]$ iff a is finite, and for all $b \in a$: $b \in HF^1[\mathcal{P}]$.

To see what this result has to do with our proposition, note that by proposition 2.9, a model is characterized up to bisimulation by a theory of $\mathcal{L}_\omega$ iff it is of rank $\leq \omega$. By proposition 2.12 it follows that the solution of a model is uniquely characterized by an $\mathcal{L}_\omega$ theory iff the model is of rank $\leq \omega$. Since the class $HF^1[\mathcal{P}]$ is exactly the class of solutions of models that are bisimilar to a finitely branching model, the result above follows.

A Hennessy-Milner class is any class of models $\mathcal{C}$ such that for all (K, w) and (K', w') in $\mathcal{C}$, $(K, w) \simeq (K', w')$ iff $(K, w) \equiv_\omega (K', w')$. A class $\mathcal{C}$ of models is a maximal Hennessy-Milner class iff for each model (K, w) that is not in $\mathcal{C}$, $\mathcal{C} \cup \{(K, w)\}$ is not a Hennessy-Milner class.

One of the results in Hollenberg (1995) is the following:

- The intersection of all maximal Hennessy-Milner classes is exactly the class of models that are bisimilar to an image-finite model.

This proposition is related to ours in the following way. First note that being 'image-finite' is another term for being 'finitely branching.' We show that the class of models of rank $\leq \omega$ is the intersection of all maximal Hennessy-Milner classes.

Assume $\mathsf{rank}(K, w) \leq \omega$. Let $\mathcal{C}$ be any maximal Hennessy-Milner class. Since $\mathsf{rank}(K, w) \leq \omega$, it holds that for each $(K', w') \in \mathcal{C}$ that $(K, w) \equiv_\omega (K', w')$ iff $(K, w) \simeq (K', w')$. So $\mathcal{C} \cup \{(K, w)\}$ is a Hennessy-Milner class, and hence $(K, w) \in \mathcal{C}$ by maximality.

Assume $\mathsf{rank}(K, w) > \omega$. Then there is a model (K', w') such that $(K, w) \equiv_\omega (K', w')$, but $(K, w) \not\simeq (K', w')$. Let $\mathcal{C}$ be a Hennessy-Milner class that contains (K', w'). Then clearly, $(K, w) \notin \mathcal{C}$, and hence (K, w) is not in the intersection of all maximal Hennessy-Milner classes.

Fagin (1995, page 232) formulates and proves the following result:

- The uniqueness ordinal of a Kripke model with at most finite fanout is at most ω. Conversely, if K is a nonflabby Kripke model with uniqueness ordinal at most ω, then K has finite fanout.

We have seen earlier how the 'uniqueness ordinal' is defined: it corresponds to our notion of rank. A Kripke-model is in the extension of the term 'nonflabby' iff for all indices w, w' in K, there is an $\mathcal{L}_\infty$-sentence ϕ such that $w \models \phi$, but $w' \not\models \phi$. The notion of 'having finite fanout' is of course the same as 'being finitely branching.'

That Fagin's proposition follows from ours can be seen by showing for the left-to-right part that each model is bisimilar to a 'nonflabby' model by simply 'dividing out' by $\simeq$, which results in a nonflabby model by proposition 2.9. For the other direction, note that if two Kripke models are bisimilar, then they have the same 'uniqueness ordinal.'

2.7 Conclusions

We have seen in this chapter how bisimulation provides us with a way of making explicit the connections between infinitary modal logic and its semantics in the form of Kripke models, knowledge structures or sets. I have studied these interrelationships in some detail with the help of bounded bisimulations, and I have shown how the tools developed here can be used to connect results from non-well-founded set-theory with results on knowledge structures and process algebra.

The close relations between Kripke models, their set-theoretic counterparts and descriptions of such models in a modal language play an important role in this dissertation. A point that will be important in several places in this dissertation is the fact that there are many 'more' Kripke models than there are possibilities or knowledge structures: each possibility corresponds with a whole class of bisimilar, but structurally different, models. In other words, a semantics for modal logic in the form of Kripke models has a finer structure than a semantics in terms of non-well-founded sets. The results of this chapter show that this fine-grainedness of Kripke semantics does not play any essential role in the semantics for (infinitary) classical modal logic: bisimilar models always satisfy the same sentences. One question that will concern us in the next chapter is whether the distinctions between bisimilar, but different, models are important for epistemic semantics at all.

3 Epistemic Logic

Epistemic logic is the logic of knowledge and belief. The topic of this chapter is multi-agent propositional epistemic logic: the logic of knowledge and belief that different agents have about the world and about the information of each other. The chapter is divided into five sections.

The first section is a short introduction to multi-modal epistemic logic. Its purpose is to introduce some basic terminology that is used in the rest of the chapter, and to present the definitions of satisfaction of sentences of epistemic logic that were given in the previous chapters from a more philosophical point of view.

The three sections following section 3.1 are each devoted to a particular operator that denotes a way that a group of agents can be said to know certain facts. In section 3.2 I discuss the notion of common knowledge. A piece of information is common knowledge in a group just in case each of the agents has the information, each of the agents knows that each of the agents has it, and so on. I will discuss different ways of defining the concept, and study the logic of the operator.

Section 3.3 is about distributed information, which is the information that would result if one were to add all information of the separate agents together. It turns out that it is particularly hard to define a more or less robust notion of distributed knowledge in a formally precise way, and a major part of the chapter is concerned with this problem. Again, we study the logic and give a completeness proof.

Distributed knowledge is an 'external' notion: it characterizes the information that another agent would have if she knew what the agents in the group know. Section 3.4 studies the notion of 'combined knowledge,' which is meant to capture the information that agents would have if they were to share their information with each other. This notion is different from that of distributed knowledge, and I will study the differences and connections in some detail.

In the last section of this chapter, I consider the concepts introduced earlier in a different light. In this section, I will give a sketch of a semantics for a language in which we can talk about information about a certain topic.

3.1 Propositional Epistemic Logic

The language of epistemic logic is given in the following definition.

Definition 3.1 (language of epistemic logic)
Given a set of agents $\mathcal{A}$, and a set of propositional variables $\mathcal{P}$, the *classical language of epistemic logic* $\mathcal{L}$ is the smallest set that includes $\mathcal{P}$ and for which it holds that if ϕ and ψ are in $\mathcal{L}$, then also $\neg\phi$, $\phi \wedge \psi$, and $\Box_a\phi$ are sentences of $\mathcal{L}$.

I will use the standard abbreviations: $\Diamond_a\phi$ for $\neg\Box_a\neg\phi$, $\phi \to \psi$ for $\neg(\phi \wedge \neg\phi)$, and $\phi \vee \psi$ for $\neg\phi \to \psi$. Also, if Φ is a (finite) set of sentences $\{\phi_1 \ldots \phi_n\}$, I will write $\bigwedge \Phi$ for $(\phi_1 \wedge (\phi_2 \wedge \ldots \phi_n))$ □

The classical language of multi-modal logic contains the operators from propositional logic, together with sentential operators of the form $\Box_a$. The intended interpretation of a sentence of the form $\Box_a\phi$ is that agent a knows or believes that ϕ. However, the terms 'agent,' 'knowledge' and 'belief' should be read in a very loose sense. An 'agent' can be any kind of object for which it makes sense to say that it has information: humans, but also robots, database systems, and, in a more abstract way, software agents or even electronic devices. Saying that human agents know and believe certain things is not a very controversial thing to do; ascribing knowledge and belief to abstract agents is more controversial. The literature on computer science and artificial intelligence contains a fairly sophisticated and well-motivated theory of agenthood (Wooldridge and Jennings (1994) give an overview), and Fagin et al. (1995) provide a systematic method for ascribing knowledge to such abstract agents using possible worlds semantics. The latter approach will be discussed in more detail in section 5.3.

One way of providing a semantics for the language of epistemic logic is in terms of Kripke models (Kripke (1963a)).

Definition 3.2 (Kripke models)
A *pointed Kripke model* (K, x) is a quadruple $(W, (\xrightarrow{a})_{a\in\mathcal{A}}, V, x)$, where W is any set, x is a distinguished element of W (the point of evaluation), $\xrightarrow{a}$ is a relation on W for each $a \in \mathcal{A}$, and V is a valuation function that assigns to each propositional variable $p \in \mathcal{P}$ a subset of W. □

The definitions of satisfaction of sentences of $\mathcal{L}$ in a model is as follows.

Definition 3.3 (truth in Kripke models)

$$\begin{array}{rcl}
(K,x) \models p & \text{iff} & x \in V(p) \\
(K,x) \models \phi \wedge \psi & \text{iff} & (K,x) \models \phi \text{ and } (K,x) \models \psi \\
(K,x) \models \neg\phi & \text{iff} & (K,x) \not\models \phi \\
(K,x) \models \Box_a\phi & \text{iff} & \text{for all } y \text{ such that } x \xrightarrow{a} y : (K,y) \models \phi
\end{array}$$

Where it does not lead to confusion, I will write x instead of (K, x). In what follows, I will often refer to the set of pointed models (K, y) such that $x \xrightarrow{a} y$ as the *information state* of a in (K, x), and sometimes write $x(a)$ for this set.

The intuition behind this semantics as a semantics for epistemic logic derives from Hintikka (1962). The idea is that we can model the belief of an agent as "the set of possibilities that are compatible with the belief of that agent," or, alternatively but not incompatibly, as "the set of possibilities that, with respect to the beliefs of the agent, could be (models of) the real world." The clause in the definition simply says that a believes that ϕ in world x just in case ϕ is true in all worlds in the information state of a in x.

A pointed Kripke model (K, x) is a representation of certain facts about a situation together with a representation of the information that a group of agents has about these facts. The facts are modeled by the values of the propositional variables, and the information of the agents is modeled by the accessibility relations. If an index y is accessible from x by an arrow labeled with a, the situation modeled by (K, y) is compatible with the information that a has in the situation (K, x). In short, the information represented by a Kripke model (K, x) is characterized by two factors: the value of the propositional variables, and the information states of each of the agents. An information state is a set of Kripke models.

If we use non-well-founded set-theory, we can implement this way of modeling the world in a more direct way:

Definition 3.4 (possibilities)
Let $\mathcal{A}$, a set of agents, and $\mathcal{P}$, a set of propositional variables, be given. The class of possibilities is the largest class such that:

- A possibility w is a function that assigns to each propositional variable $p \in \mathcal{P}$ a truth value $w(p) \in \{0, 1\}$ and to each agent $a \in \mathcal{A}$ an information state $w(a)$.
- An information state σ is a set of possibilities. □

A possibility w characterizes which propositions are true and which are false, and it characterizes the information of each of the agents in the form of an information state σ, that consists of the set of possibilities the agent considers possible in w.

This definition of possibilities ranges over the universe of non-well-founded sets. The formal aspects are explained in more detail in chapter 1.

Truth of classical modal sentences in a possibility can be defined in a way analogous to the definition of truth in Kripke models.

Definition 3.5 (truth in possibilities)
Let w be a possibility.

$$w \models p \quad \text{iff} \quad w(p) = 1$$

$$\begin{array}{rll} w \models \phi \wedge \psi & \text{iff} & w \models \phi \text{ and } w \models \psi \\ w \models \neg\phi & \text{iff} & w \not\models \phi \\ w \models \Box_a \phi & \text{iff} & \text{for all } v \in w(a) : v \models \phi \end{array}$$ □

We saw in the previous chapter that there is a close relation between possibilities and pointed Kripke models.

Definition 3.6 Let $(K, x) = (W, (\xrightarrow{a})_{a \in \mathcal{A}}, V, x)$ be a pointed Kripke model.

- A *decoration* of (K, x) is a function δ that assigns to each world $y \in W$ a function δ_y with $\mathcal{P} \cup \mathcal{A}$ as its domain, such that $\delta_y(p) = 1$ iff $y \in V(p)$ for each $p \in \mathcal{P}$, and $\delta_y(a) = \{\delta_z \mid y \xrightarrow{a} z\}$ for each $a \in \mathcal{A}$.

- If δ is a decoration of (K, x), then δ_x is called its *solution*, and (K, x) is a *picture* of δ_x. □

A decoration of a Kripke model assigns to each possible world x in the model a possibility that assigns the same truth-values to the propositional variables as they get in the model at x, and that assigns to each agent a the set of possibilities that are assigned to worlds accessible from x by $\xrightarrow{a}$.

The notions of solution and picture give us a correspondence between Kripke-models and possibilities:

Proposition 3.7 (cf. chapter 2)

- Each Kripke model has a unique decoration, and therefore a unique solution. Moreover, the solution of a Kripke model is a possibility.

- Each possibility has a Kripke model as its picture.

- Two Kripke-models are pictures of the same possibility iff they are bisimilar. □

Defining truth of a formula in a Kripke model in the standard way, it holds that:

Proposition 3.8 (cf. chapter 2) For each possibility w:

$$w \models \phi \text{ iff } \phi \text{ is true in each picture of } w$$ □

So a possibility and a picture of it satisfy exactly the same sentences. This means that one can see possibilities as representatives of equivalence classes of Kripke models under bisimulation.

In this dissertation, possibilities will be used as the primary model for epistemic logic, although I will often refer to Kripke models as well. We have seen that the use of possibilities means that certain distinctions between Kripke models collapse: any two different but bisimilar Kripke models correspond to the same possibility. This suggests a view on Kripke models in which the differences between bisimilar Kripke models is somehow not important.

If Kripke models are used as transition systems to represent the behavior of processes, the differences between bisimilar Kripke models are often ignored. if two processes are represented by bisimilar graphs, then these processes cannot be distinguished by the actions they perform: they are 'observationally equivalent.' So, when we are interested in the behavior of processes only, it is warranted (and perhaps even desirable) to use possibilities instead of Kripke models. In the last chapter of Aczel (1988) such a model is developed.

The question whether this is warranted for epistemic semantics as well is difficult, because there is no real consensus about what an index in a Kripke model is meant to represent. I will limit the discussion of this question to some preliminary remarks here, and return to the topic later when we discuss the notion of distributed knowledge.

We saw in chapter 2 that different models for epistemic logic (Kripke models, possibilities and knowledge structures) are all equivalent 'modulo bisimulation.' Also, classical (infinitary) epistemic logic cannot distinguish between bisimilar models. The differences between bisimilar Kripke models can be, and usually are, ignored in a lot of the formal work on epistemic semantics; in completeness proofs, for example, models are being unraveled and quotients are taken of them without any further discussion. All this is circumstantial evidence for the correctness of ignoring distinctions between bisimilar Kripke models and using possibilities as our models instead.

I believe that when one takes the idea of Kripke models as *models* very seriously, one can make as serious argument for collapsing distinctions between bisimilar models. Roughly, Kripke models are models of some real or imaginary situation that have a certain idealized structure in common with these real or imaginary situations. A world in a Kripke model *represents* a state of affairs, and it does so in a very limited way: it only tells you of certain facts whether they obtain or not and it tells you what different agents believe. If two sets σ and τ of pointed models each represent the same sets of possible situations, then σ and τ represent the same information state. That means that two sets of indices in a Kripke model represent the same beliefs just in case for each index in the first set, there is an index in the second set that represents the same situation, and vice versa. These two aspects of the model, truth values and information states, are all that is important about the model.

We can represent this view in a single (rather long) slogan:

> Two pointed Kripke models w and v represent the same situation just in case the propositional variables get the same interpretation, and if an index is a-accessible from w in K, then there must be an index that represents the same situation and that is a-accessible from v in K. And vice versa.

This is just the definition of bisimulation.

Logical Omniscience

Possible world semantics gives rise to a rather distorted notion of belief and knowledge. In particular, it holds that if ψ is a logical consequence of ϕ, then in all models in which $\Box_a\phi$ is true (i.e. where agent a knows that ϕ is true), $\Box_a\psi$ is true as well. In other words, the belief (and the knowledge) of an agent is closed under logical consequence. This does not conform at all to the situation as it actually is: people do not always see the consequences of their beliefs, and know that other people are limited in the same way. This fact is particularly apparent when we look at mathematical truths: once we know the basic axioms of arithmetic, we do not know that Fermat's last theorem is true, although it is a logical consequence of the axioms. This mismatch between the semantics and the concept of knowledge and belief is known as "the problem of logical omniscience."

If epistemic semantics is seen as a proposal for a theory describing linguistic usage of phrases such as "John knows that ...," then the problem of logical omniscience clearly falsifies the possible worlds analysis of knowledge. "Bill knows that Fermat's theorem is true" is a contingent sentence. The corresponding logical formula is a tautology.

This does not imply that the possible worlds analysis of knowledge and belief is useless as a model in which we can study aspects of knowledge and belief. Modeling something implies simplification and abstraction from many of the 'inessential' properties of whatever it is that is modeled. Whether the fact that agents are logically omniscient is decisive for the failure of the model to be a model of knowledge depends on whether the lack of logical omniscience in reality is taken to be a very important aspect of our concept of knowledge or not. Even if the operators $\Box_a$ are not a perfect reflection of the notions of knowledge or belief, they can still be of interest for a theory of the more abstract concept of *information*. This is what Barwise (1989) proposes, in a slightly different setting: to use the term 'information' instead of 'knowledge' as the informal notion that best corresponds to the interpretation of $\Box_a$. In contrast with knowledge and belief, it does make sense to say that *information* is closed under logical consequence: if a sentence logically follows from the information you already have, then in some sense, you also have the information that this sentence is true, even though you may not be aware of it.

In this dissertation, I will use the terms 'belief,' 'knowledge' and 'information' more or less interchangeably. It is perhaps best to read the verb 'to believe' as short for 'to have the information that' and 'to know' as short for 'to have the true information that.'

Introspection and Factivity

It is often taken to be the case that knowledge and belief are *introspective*: if you believe something, then you believe that you believe it, and if you don't believe something, then you believe that you don't believe it.

Another property that is important in epistemic logic is the notion of *factivity*. Factive verbs imply that their complement is true. Knowledge is factive, belief is not: John may believe things that are false, but if he *knows* something, then it must be true.[1]

We can express these constraints on information in our formal language in the form of the following axiom schemes:

positive introspection $\Box_a \phi \rightarrow \Box_a \Box_a \phi$

negative introspection $\neg \Box_a \phi \rightarrow \Box_a \neg \Box_a \phi$

full introspection $\neg \Box_a \phi \rightarrow \Box_a \neg \Box_a \phi$ and $\Box_a \phi \rightarrow \Box_a \Box_a \phi$

factivity $\Box_a \phi \rightarrow \phi$

There is a very close connection between these axiom schemes and constraints on possibilities. If the information of an agent is introspective, this means that in each of her epistemic alternatives, she will be in the information state she is actually in. In our formal semantics, this means that if an agent a is in a state σ, then for each possibility $v \in \sigma$ it holds that $v(a) = \sigma$. Similarly, if the information of an agent in factive, then it is consistent with the facts as they are in the actual world: the actual world is among the epistemic alternatives of the agent. Formally, this means that if the information of an agent a in a possibility w is factive then $w \in w(a)$.

Consider now the following classes of possibilities:

Definition 3.9 A class of possibilities S is *closed* iff it holds that if $w \in S$ and $v \in w(a)$ then $v \in S$.

1. The class of positively introspective possibilities $\mathcal{P}$ is the largest closed class such that for each $w \in \mathcal{P}$ it holds that $v \in w(a)$ implies that $v(a) \subseteq w(a)$
2. The class of negatively introspective possibilities $\mathcal{N}$ is the largest closed class such that $w \in \mathcal{N}$ and $v \in w(a)$ imply $w(a) \subseteq v(a)$
3. The class $\mathcal{I}$ of fully introspective possibilities is the (closed) class $\mathcal{P} \cap \mathcal{N}$.
4. The class of reflexive possibilities $\mathcal{T}$ is the largest closed class such that $w \in \mathcal{T}$ implies $w \in w(a)$ $\Box$

[1] But see Lenzen (1978), who argues that the assumption that knowledge and belief have these properties leads to counterintuitive predictions about the interaction between the two operators.

I will often refer to possibilities that are fully introspective as *belief models*, and to possibilities that are introspective and reflexive as *knowledge models*.

In chapter 2, we saw that possibilities are intimately connected with the language of infinitary modal logic. In fact, if we consider infinitary modal logic, the axiom schemes characterize the corresponding class of possibilities exactly:

Proposition 3.10 Let $\mathcal{L}_\infty$ be the language of infinitary modal logic (cf. definition 2.5). Then:

$$\begin{array}{rcl} \forall v \in w(a) \Rightarrow v(a) \subseteq w(a) & \text{iff} & w \models \Box_a\phi \to \Box_a\Box_a\phi \text{ for each } \phi \in \mathcal{L}_\infty \\ \forall v \in w(a) \Rightarrow w(a) \subseteq v(a) & \text{iff} & w \models \neg\Box_a\phi \to \Box_a\neg\Box_a\phi \text{ for each } \phi \in \mathcal{L}_\infty \\ w \in w(a) & \text{iff} & w \models \Box_a\phi \text{ for each } \phi \in \mathcal{L}_\infty \end{array}$$

proof: The proof is simple with the result of corollary 2.23 that each possibility can be characterized by a single sentence of $\mathcal{L}_\infty$.

I'll consider two cases: factivity and negative introspection. The proof for positive introspection is completely similar.

To see that $\neg\Box_a\phi \to \Box_a\neg\Box_a\phi$ is true in any negatively introspective possibility, assume that w is negatively introspective and that $w \models \neg\Box_a\phi$. Then there must be a $u \in w(a)$ such that $u \models \neg\phi$. Now take any $v \in w(a)$. Since $w(a) \subseteq v(a)$, it follows that $u \in v(a)$, so $v \models \neg\Box_a\phi$. Since v was arbitrary, we have that $w \models \Box_a\neg\Box_a\phi$.

The interesting case is the other direction, for which we need to show that any possibility w in which all instantiations of the axiom $\neg\Box_a\phi \to \Box_a\neg\Box_a\phi$ are true is positively introspective. Fix some $v \in w(a)$, and take any u such that $u \in w(a)$. We know by proposition 2.23 that for each possibility there is a sentence of infinitary modal logic that is true only in that possibility. Let ϕ_u be a sentence that is true at u only. Since $u \in w(a)$, it holds that $w \models \neg\Box_a\neg\phi_u$. With the axiom for negative introspection, we conclude that $w \models \Box_a\neg\Box_a\neg\phi_u$. In particular, $v \models \neg\Box_a\neg\phi_u$. But then, there must be a $u' \in v(u)$ such that $u' \models \phi_u$, and since ϕ_u is true at u only, it follows that $u' = u \in v(a)$.

To show that reflexivity is characterized by the infinitary axiom scheme $\Box_a\phi \to \phi$, suppose first that $w \in w(a)$, and that $w \models \Box_a\phi$. Then, $v \models \phi$ for each $v \in w(a)$, and, since also $w \in w(a)$, we have that $w \models \phi$.

For the other direction, assume that $w \models \Box_a\phi \to \phi$ for each ϕ, and assume to the contrary that $w \notin w(a)$. We know that there is a sentence ϕ_w that is true only at w. So, we can infer that $w \models \Box_a\neg\phi_w$. But then, $w \models \neg\phi_w$, a contradiction. □

The relation between properties of Kripke models and the axiom schemes of epistemic logic have been studied well. The four axioms above each correspond to a particular property of 'frames.'[2] Each of the classes of possibilities given

[2]More precisely: a *frame* is a Kripke model without a valuation function; we build a Kripke model from a frame by adding a valuation function. We say that a sentence is valid in a frame iff

above contains exactly the possibilities that are solutions of Kripke models that have the corresponding frame property. For example, each member of $\mathcal{P}$ has a transitive picture, and each transitive Kripke model has a member of $\mathcal{P}$ as its solution.

I have summarized the results in figure 3.1.

Class of possibilities	Property of possibility w	Axiom Scheme
Positive Introspection ($\mathcal{P}$)	$v \in w(a) \Rightarrow v(a) \subseteq w(a)$	$\Box_a\phi \to \Box_a\Box_a\phi$
Negative Introspection ($\mathcal{N}$)	$v \in w(a) \Rightarrow w(a) \subseteq v(a)$	$\neg\Box_a\phi \to \Box_a\neg\Box_a\phi$
Full Introspection ($\mathcal{P} \cap \mathcal{N}$)	$v \in w(a) \Rightarrow v(a) = w(a)$	$\Box_a\phi \to \Box_a\Box_a\phi$ &
		$\neg\Box_a\phi \to \Box_a\neg\Box_a\phi$
Reflexivity ($\mathcal{T}$)	$w \in w(a)$	$\Box_a\phi \to \phi$

Figure 3.1. Properties of Possibilities

Logic

I have discussed the semantics of the language of epistemic logic; I will now discuss the logic that the semantics gives rise to.

Definition 3.11 (validity)
A sentence ϕ is valid, written as $\models \phi$, iff ϕ is true in all possibilities.
$\models_{\mathsf{belief}} \phi$ iff ϕ is true in all introspective possibilities.
$\models_{\mathsf{knowledge}} \phi$ iff ϕ is true in all introspective reflexive possibilities. □

The three notions of validity can be axiomatized:

Definition 3.12 (axioms for modal logic)
The logic K is given by the following set of axioms and rules:

axioms

A1 $\vdash \phi$ if ϕ is a propositional tautology.

A2 $\vdash \Box_a(\phi \to \psi) \to (\Box_a\phi \to \Box_a\psi)$

rules

MP If $\vdash \phi$ and $\vdash \phi \to \psi$, then $\vdash \psi$.

it is true in all worlds in Kripke models that can be build from that frame. It turns out that all instantiations of positive introspection are valid in exactly all frames in which the accessibility relations are transitive (if $w \xrightarrow{a} v \xrightarrow{a} u$, then $w \xrightarrow{a} u$; that negative introspection corresponds to euclidity (if $w \xrightarrow{a} v$ and $w \xrightarrow{a} u$, then $v \xrightarrow{a} u$), and that factivity corresponds to reflexivity ($w \xrightarrow{a} w$).

Nec If $\vdash \phi$, then $\vdash \Box_a \phi$.

The logic $\mathsf{K45}$ is as K, but with the axioms of positive and negative introspection added, and the logic $\mathsf{S5}$ is $\mathsf{K45}$ together with the factivity axiom (the logics have these names for historical reasons). We write $\vdash \phi$ or $\vdash_{\mathsf{K}} \phi$ if ϕ is derivable in K, and write $\vdash_{\mathsf{K45}}$ and $\vdash_{\mathsf{S5}}$ for derivability in the other two systems. Also, if Σ is a set of sentences, we write $\Sigma \vdash \phi$ if there are $\phi_0 \dots \phi_n \in \Sigma$ such that $\vdash \bigwedge_{i \leq n} \phi_i \rightarrow \phi$. $\Box$

It is well-known that the logic K is sound and complete with respect to validity in the class of all Kripke models, and that $\mathsf{K45}$ (and $\mathsf{S5}$) are sound and complete with respect to validity in all transitive and euclidean (and reflexive) Kripke models. The corresponding results hold for the non-well-founded semantics:

Proposition 3.13 (soundness and completeness)

$$\begin{array}{rcl} \vdash \psi & \text{iff} & \models \psi \\ \vdash_{\mathsf{K45}} \psi & \text{iff} & \models_{\mathsf{belief}} \psi \\ \vdash_{\mathsf{S5}} \psi & \text{iff} & \models_{\mathsf{knowledge}} \psi \end{array}$$

proof: Proving soundness, that each derivable sentence is always true, is a simple matter of checking each axiom and rule: we have done two cases in the proof of proposition 3.10.

The converse, completeness, follows from the fact that the logics are complete with respect to validity in the corresponding class of Kripke models. Combining this with the fact that each of these models have solutions in which the same sentences are true, and that these solutions are introspective and reflexive when the corresponding models are, gives us completeness.

We can also give a more direct argument: the standard completeness proof for Kripke models is easily adapted to apply to possibilities. We say that a set of sentences Σ is consistent if $\Sigma \not\vdash \bot$, and maximally consistent if adding any sentence to Σ results in an inconsistent set. For each maximal consistent set of sentences Σ, we define the canonical possibility w_Σ as:

$$\begin{array}{rcl} w_\Sigma(p) & = & 1 \text{ iff } p \in \Sigma \\ w_\Sigma(a) & = & \{w_\Gamma \mid \Gamma \text{ is maximal consistent and } \Box_a\phi \in \Sigma \Rightarrow \phi \in \Gamma\} \end{array}$$

It is easy to show, by induction on the structure of ϕ, that $w_\Sigma \models \phi$ iff $\phi \in \Sigma$. Also, it can be shown that if we restrict the construction by using only sets that are maximally consistent in $\mathsf{K45}$ ($\mathsf{S5}$), then the possibility w_Σ is in the class of introspective (introspective and reflexive) possibilities.

To see that these results provide us with a completeness proof, assume that a certain sentence ϕ is not derivable. Then, there is a maximal consistent set Σ that has $\neg\phi$ as a member, and then $w_\Sigma \not\models \phi$. So, ϕ is not valid. $\Box$

3.2 Common Knowledge

A sentence ϕ is common knowledge in a group of agents $\mathcal{B}$ just in case each agent in $\mathcal{B}$ knows ϕ, each agent knows of each other agent that he knows ϕ, each agent knows that each agent knows that each agent knows that ϕ, etcetera, *ad infinitum.* Apart from the term 'common knowledge,' one also finds the terms 'mutual belief' and 'joint knowledge' in the literature.[3] The concept was first introduced by Lewis (1969) in his analysis of convention, and has been used in the literature on linguistic pragmatics (e.g. Lewis (1979) and Clark and Marshall (1981)), in game theory (Aumann (1976)), and in the literature on distributed systems in computer science (Halpern and Moses (1990), Fagin et al. (1991)).

There have been different formal definitions and characterizations of common knowledge. Barwise (1989) compares three characterizations of common knowledge, and concludes that in situation theory, all three give different results. In the 'iterated approach,' common knowledge is defined in the way I introduced it above: a sentence is common knowledge just in case each agent believes the sentence, each agent believes that each of the other agents believes that sentence, etcetera. This approach has been criticised on the ground that each of the agents would have to know an infinite number of logically independent facts for a sentence to be common knowledge, which is considered implausible.

The other two approaches do not suffer from this problem. In the 'fixed point approach,' common knowledge is defined as a self-referential concept. In order for a sentence ϕ to be common knowledge in the fixed point approach, each agent has to know that ϕ is true, and moreover, each of the agents has to know that ϕ is common knowledge. Knowledge of the 'infinite number' of facts in of the iterated approach then follows from the agents just knowing these two things.

The third approach discussed by Barwise is the so-called 'shared situation approach:' here, common knowledge is defined as knowledge that a certain situation obtains that is 'shared' by each of the agents. If there is a situation that gives all agents reason to assume that each agent knows that ϕ, and all agents have sufficient reason to assume that each agent knows that this very situation in fact obtains, then the agents may conclude that it is common knowledge that ϕ. This kind of definition has been put forward by Lewis (1969), Schiffer (1972) and Clark and Marshall (1981). A typical example of a shared situation would be the scoreboard during a baseball game. Each of the participants in the game have access to the information displayed on the scoreboard, and all participants know that everyone can see the scoreboard. Another example is the utterance of a sentence followed by an acknowledgment of the hearer: such a sequence of

[3]The results of Halpern and Moses (1990) in the context of message-passing systems show that if one reads the 'knowledge' in 'common knowledge' in the strong sense as implying truth, it can never happen that any non-trivial information becomes common knowledge; at least not under the quite reasonable assumptions that message passing takes time, and is never completely reliable.

events would be reason enough to assume that the fact that the utterance is made is now mutually believed.

The three notions can be, roughly, represented as follows:

The *iterated approach* is just a straightforward rewriting of the informal definition above. Writing $C^{\mathsf{iter}}\phi$ for ϕ is common knowledge under the iterated approach, we can define:

$$w \models C^{\mathsf{iter}}\phi \quad \text{iff} \quad w \models \Box_{a_1} \ldots \Box_{a_n}\phi \quad \text{for each sequence } a_1 \ldots a_n \text{ of agents in } \mathcal{A}$$

In the *fixed point approach* a sentence ϕ is common knowledge just in case each agent knows that ϕ and each agent knows that ϕ is common knowledge. Given this, we can characterize common knowledge of ϕ as that property P of possibilities that holds of a possibility w just in case it holds in w that a knows that ϕ and that a knows that the property P holds. If we denote this property by '$\models C^{\mathsf{fix}}\phi$' we can formally express this by the following equivalence:

$$w \models C^{\mathsf{fix}}\phi \quad \text{iff} \quad \forall a \forall v \in w(a) : v \models \phi \text{ and } v \models C^{\mathsf{fix}}\phi$$

Since this does not uniquely identify a property, we let $\models C^{\mathsf{fix}}\phi$ be the *largest* property that satisfies the equation above.[4]

In the *shared situation approach*, a sentence ϕ is mutually believed just in case there is a situation σ in which (1) ϕ holds, and (2) the situation σ implies, or gives reason enough to assume, that each of the agents knows (or believes) that the situation σ in fact obtains, and (3) each of the agents does believe that σ obtains.

If we identify a situation with a set of possibilities ('all possibilities that represent that situation', or, more situation-theoretically, 'all maximal extensions' of that situation') we can transpose Barwise's analysis to our framework and define:

$$\begin{array}{lll} w \models C^{\mathsf{share}}\phi & \text{iff} & \text{there is a set of possibilities } \sigma \text{ such that:} \\ & & (1)\ v \in \sigma \Rightarrow v \models \phi \\ & & (2)\ v \in \sigma \Rightarrow v(a) \subseteq \sigma \text{ for each } a \\ & & (3)\ w(a) \subseteq \sigma \text{ for each } a \end{array}$$

If we compare the three definitions, it turns out that all three are equivalent:

Proposition 3.14 For each possibility w:[5]

$$w \models C^{\mathsf{iter}}\phi \;\Leftrightarrow\; w \models C^{\mathsf{fix}}\phi \;\Leftrightarrow\; w \models C^{\mathsf{share}}\phi$$

[4] Alternatively, and perhaps more clearly, we can define the class of possibilities in which C^{fix} is true as the largest closed class (cf. definition 3.9) such that for any possibility w in the class, and any agent a, $w \models \Box_a\phi$.

[5] Fagin et al. (1995) prove that the iterated and the fixed point definitions are equivalent in Kripke models.

proof:

[From the iterated account to the fixed points] Assume $w \models C^{\mathsf{iter}}\phi$. It is not hard to see that it holds that: $w \models C^{\mathsf{iter}}\phi$ iff $\forall a \forall v \in w(a)$: $v \models \phi$ and $v \models C^{\mathsf{iter}}\phi$

Since we have defined $\models C^{\mathsf{fix}}\phi$ as the largest set with exactly this property, it follows that $w \models C^{\mathsf{fix}}\phi$.

[From fixed points to shared situations] Consider the set $\sigma = \{v \mid v \models \phi$ and $v \models C^{\mathsf{fix}}\phi\}$. Then $v \in \sigma$ implies that $v(a) \subseteq \sigma$ by definition of C^{fix}, and clearly, $v \models \phi$ for each $v \in \sigma$. Assume $w \models C^{\mathsf{fix}}\phi$. Then clearly, $w(a) \subseteq \sigma$, so $w \models C^{\mathsf{share}}\phi$.

[From shared situations to the iterated approach] Assume $w \models C^{\mathsf{share}}\phi$. We need to show that $w \models \Box_{a_1} \dots \Box_{a_n}\phi$ for each sequence $a_1 \dots a_n$ of agents. That this holds is easily proven by an induction on n. $\Box$

Thus, in a classical possible worlds framework (the definitions can be easily reformulated to apply to Kripke models, and the equivalence results will continue to hold) the three different characterizations of common knowledge collapse.

I am not sure whether this result should be seen as a positive or a negative one. In contrast to the analysis above, on Barwise's analysis in situation theory of the three definitions, all three turn out to give different situation-theoretic notions of common knowledge. Depending on one's intuitions about the three notions, this might be seen both as a weakness of classical logic (classical logic cannot distinguish between intuitively different notions) or as a weakness of the situation-theoretic approach (situation theory makes unnatural distinctions between intuitively equivalent notions). Of course, there are important differences between the three definitions, in the sense that they derive from distinct intuitions about what common knowledge is, or how it arises. We cannot model such 'intensional' differences with the use of possibilities, but neither can situation theory. The differences between the three kinds of definitions in the analysis of Barwise are only visible at the transfinite level; restricting Barwise's analysis to models in which agents believe only finitely logically independent facts (which is a natural assumption), the three notions collapse also in situation theory. This makes it, at least to me, very hard to see how the distinctions between the three characterizations correspond to pre-situation-theoretic distinctions. To put it bluntly: it seems that situation theory is making trouble where there was no trouble to be found.

Whatever the conclusion is, the fact that the three different characterizations come down to the same semantical characterization in our framework makes the choice between the definitions meaningless: we can take either one.

Expressive Power

We now turn to the logic of the operator C. Instead of considering a language with a single modal operator C, I will study a slightly richer language, containing an operator $C_{\mathcal{B}}$ for each set of agents $\mathcal{B}$. I will use the symbol $\mathcal{L}^C$ refer to the language of classical modal logic extended with the operators $C_{\mathcal{B}}$. The intended interpretation of $C_{\mathcal{B}}\phi$ is that ϕ is common knowledge among the agents in $\mathcal{B}$. The semantics of the new operators is exactly what one would expect. I give the fixed point definition here:

Definition 3.15 A class of possibilities S is $\mathcal{B}$-*closed* iff for each $v \in S$, it holds that $v(b) \subseteq S$ for each $b \in \mathcal{B}$.

We define *the common knowledge in* w, $C_{\mathcal{B}}(w)$ as the smallest $\mathcal{B}$-closed class S such that $w(a) \subseteq s$ for each $a \in \mathcal{B}$. The sentence $C_{\mathcal{B}}\phi$ is true in w iff ϕ is true in all possible worlds in $C_{\mathcal{B}}(w)$:

$$w \models C_{\mathcal{B}}\phi \quad \text{iff} \quad v \models \phi \text{ for each } v \in C_{\mathcal{B}}(w)$$ □

The language $\mathcal{L}^C$ is richer in expressive power than $\mathcal{L}$: there are sentences in $\mathcal{L}^C$ not equivalent to any sentence of $\mathcal{L}$. In fact, we can show that there are sentences in $\mathcal{L}^C$ that cannot even be characterized by any *set* of sentences of $\mathcal{L}$. To see this, let the possibilities w_k be given by: $w_0(p) = 0$, $w_0(a) = \emptyset$, $w_{n+1}(p) = 1$, $w_{n+1}(a) = \{w_n\}$. We let $w_\Omega(p) = 1$ and $w_\Omega(a) = \{w_\Omega\}$. Compare now the possibilities w and v such that $w(a) = \{w_k \mid k \leq \omega\}$, and $v(a) = w(a) \cup \{w_\Omega\}$. Figure 3.2 is a picture of w and v. It holds that w and v are ω-bisimilar (note

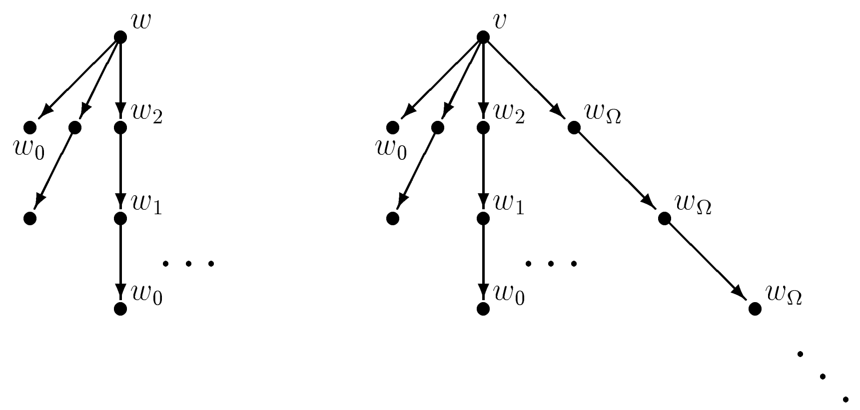


Figure 3.2.

that each w_{k+1} is k-bisimilar to w_Ω), so the same sentences of $\mathcal{L}$ are true in both possibilities. But $w \not\models \Diamond_a C_{\{a\}}p$ and $v \models \Diamond_a C_{\{a\}}p$.

Do we really need this much expressive power? Fagin et al. (1991) give an example to illustrate that one needs knowledge structures of length $> \omega$ to model certain situations; this example can also serve as an example that shows that we need a language stronger than classical modal logic to describe certain situations.

Consider a situation with two agents, a and b. We consider the case where we model the *knowledge* of the agents. The agents communicate according to the following protocol: "If you get a message from someone, send an acknowledgement that you got the message to the sender." Unfortunately, the 'channel' through which they are communicating is unreliable: they cannot be sure whether their messages have arrived (until they get an acknowledgment from the other).

In the initial situation, a knows that p, but b does not, and a sends b the message that p. So, after a tells b that p is the case, and the message arrives (it may not, since the channel is not reliable), then b knows that p and b knows that a knows that p. Following the protocol, b tries to send an acknowledgement to a that he got the message. If a gets the acknowledgement, she knows that b knows that p. Moreover, following the protocol, she sends an acknowledgement to b that she got the acknowledgement. If b receives the message, he will know that a knows that b knows that p. Since the channel is unreliable, the process may stop after any point k. Note that after exactly k rounds of successful communication, it holds that $E^k_{\{a,b\}}p$, but not $E^{k+1}_{\{a,b\}}$.[6]

Consider now the information of a third agent c who knows that a and b are communicating in the way described. Agent c cannot know how many messages have been successfully sent (the communication channel may have broken down after any finite number of rounds). So, for each k, it holds that $\Diamond_c E^k_{\{a,b\}}p$. Since c knows that only a finite number of messages have been sent, he also knows that it is not common knowledge among a and b that p is the case; it holds that $\Box_c \neg C_{\{a,b\}}p$.

If we consider the possibility that is the natural representation of this situation (and make some extra assumptions that a and b have trivial believes about c's information state), then we can show that there is a possibility in which the same sentences of $\mathcal{L}$ are true, but where $\Box_c \neg C_{\{a,b\}}p$ is false. If we want to capture distinctions such as these in our object language, a language with at least the expressive power of $\mathcal{L}^C$ is needed.

Logic

The set of sentences of $\mathcal{L}^C$ that are valid is relatively easily axiomatized. We only need to add one axiom and one rule to get a complete axiomatization for the logics of knowledge and belief.

Definition 3.16 (axioms for $C_\mathcal{B}$)

C1 $C_\mathcal{B}\phi \to \Box_a(\phi \wedge C_\mathcal{B}\phi)$ for each $a \in \mathcal{B}$.

RC If $\vdash \phi \to \Box_a(\phi \wedge \psi)$ for each $a \in \mathcal{B}$, then $\vdash \phi \to C_\mathcal{B}\psi$.

The deduction systems CK (CK45, CS5) consist of all axioms of K (K45, S5) together with this axiom and rule. □

[6]Fagin et al. (1991) use the notation $E_\mathcal{B}\phi$ as an abbreviation of the conjunction of all sentences $\Box_b\phi$ for $b \in \mathcal{B}$. $E^k_\mathcal{B}$ stands for an row of k occurrences of $E_\mathcal{B}$.

Intuitively, the axiom says that if a sentence ϕ is common knowledge in a group of agents $\mathcal{B}$, then each of the agents in $\mathcal{B}$ knows that ϕ is the case, and knows that ϕ is common knowledge.

The intuition behind the rule is the following. Assume that ϕ is such that it logically implies that each of the agents in a group $\mathcal{B}$ know that ϕ and that ψ. Since the knowledge of each of the agents is closed under logical consequence, it holds that if ϕ is true, each of the agents in $\mathcal{B}$ knows of each of the agents in $\mathcal{B}$ that they know that both ϕ and ψ is true. Since this argument can be extended *ad infinitum*, it follows that if ϕ is true, each of the agents in $\mathcal{B}$ knows that ψ, each of the agents knows of each of the other agents that he or she knows that ψ is true, etcetera. In other words, if ϕ is true, then ψ (and ϕ) is common knowledge in $\mathcal{B}$.

Instead of using the induction rule RC, we could also have introduced an additional induction axiom, namely: $\bigwedge_{a\in\mathcal{B}}(\Box_a\phi\wedge C_{\mathcal{B}}(\phi\rightarrow\Box_a\phi))\rightarrow C_{\mathcal{B}}\phi$, together with a necessitation rule for $C_{\mathcal{B}}$. The choice is a matter of taste: the two systems derive the same tautologies.[7]

Proposition 3.17 (soundness and completeness.)
The logics CK, CK45 and CS5 are sound and complete with respect to $\models$, $\models_{\mathsf{belief}}$ and $\models_{\mathsf{knowledge}}$ respectively.

proof: The basic idea for the completeness proof for CK comes from a completeness proof for propositional dynamic logic by Kozen and Parikh (1981), which is a language with more expressive power than $\mathcal{L}^C$.

The proof of Kozen and Parikh works just as well to prove completeness of CK, and, conveniently, also for CS5 and CK45. A worked-out proof for the Kripke semantics can be found in Fagin et al. (1995).

I will give a sketch of the proof using possibilities. First, we say that a set of sentences Φ is *Fischer-Ladner closed* iff Φ is closed under subformula's and if $C_{\mathcal{B}}\phi\in\Phi$, then $\Box_a C_{\mathcal{B}}\phi\in\Phi$ and $\Box_a\phi\in\Phi$. If Φ is Fischer-Ladner closed, we say that a set of sentences Σ is maximal consistent in Φ just in case Σ is consistent, and for each sentence $\phi\in\Phi$, either $\phi\in\Sigma$ or $\neg\phi\in\Sigma$.

---

[7] To see that the rule RC can be derived from this axiom and necessitation, suppose $\vdash\phi\rightarrow\Box_a(\phi\wedge\psi)$ for each a. Then, *a forteriori*, $\vdash(\phi\wedge\psi)\rightarrow\Box_a(\phi\wedge\psi)$. Using the necessitation rule we can derive $\vdash C_{\mathcal{B}}((\psi\wedge\phi)\rightarrow\Box_a(\phi\wedge\psi))$, and so, $\vdash\phi\rightarrow C_{\mathcal{B}}((\phi\wedge\psi)\rightarrow\Box_a(\phi\wedge\psi))$. We already had that $\vdash\phi\rightarrow\Box_a(\phi\wedge\psi)$ for each $a\in\mathcal{B}$, so we know that ϕ implies the two antecedents of the induction axiom. We conclude that $\vdash\phi\rightarrow C_{\mathcal{B}}(\phi\wedge\psi)$ and, since $C_{\mathcal{B}}$ distributes over conjunction, we conclude that $\vdash\phi\rightarrow C_{\mathcal{B}}\psi$.

To see that necessitation and the induction rule can be derived, suppose first that $\vdash\phi$. Then, for each $a\in\mathcal{B}$, $\vdash\Box_a\phi$ by necessitation for $\Box$, so $\vdash\phi\rightarrow\Box_a\phi$, and by the induction rule RC, $\vdash\phi\rightarrow C_{\mathcal{B}}\phi$. Applying modus ponens yields that $\vdash C_{\mathcal{B}}\phi$. To show that the induction axiom can be derived, let χ be the sentence $\bigwedge_{a\in\mathcal{B}}(\Box_a\phi\wedge C_{\mathcal{B}}(\phi\rightarrow\Box_a\phi))$. With axiom C1, it follows that $\chi\rightarrow\Box_a(\phi\wedge\chi)$ for each $a\in\mathcal{B}$ and with the induction rule we conclude that $\chi\rightarrow C_{\mathcal{B}}\phi$, which is the induction axiom.

Given a Fischer-Ladner closed set Φ, we can construct canonical possibilities w_Σ in exactly the same way as we did in the proof of proposition 3.13, except that we only use sets in the construction that are maximal consistent in Φ. We can then prove a truth-lemma: if Φ is finite and Σ is maximal consistent in Φ, then for all $\phi \in \Phi$:

$$w_\Sigma \models \phi \text{ iff } \phi \in \Sigma$$

The proof is by induction on ϕ. The only difficult part of the proof is where ϕ is of the form $C_\mathcal{B}\phi$.

First suppose that $C_\mathcal{B}\phi \in \Sigma$. Consider the set $S = \{w_\Gamma \mid C_\mathcal{B}\phi \in \Gamma\}$, and take any $w_\Gamma \in S$, and any $a \in \mathcal{B}$. Since $C_\mathcal{B}\phi \in \Gamma$ and Γ is maximal in Φ, and Φ is Fischer-Ladner closed, it follows that (i) $\Box_a C_\mathcal{B}\phi \in \Gamma$, and (ii) $\Box_a\phi \in \Gamma$. From (i), it follows by definition of w_Γ that $C_\mathcal{B}\phi \in w_\Delta$ for each $w_\Delta \in w_\Gamma(a)$. From (ii), it follows by a standard argument that $w_\Gamma \models \Box_a\phi$. Since Γ and a were arbitrary, we can conclude that for any $w \in S$ and any $a \in \mathcal{B}$, $w \models \Box_a\phi$ and $w(a) \subseteq S$. This means that S is $\mathcal{B}$-closed.

In particular, this means that for all w in the *smallest* $\mathcal{B}$-closed set that contains w_Σ, $w \models \Box_a\phi$. But then, in the smallest $\mathcal{B}$-closed set that contains $w_\Sigma(a)$ for each $a \in \mathcal{B}$, it holds that $w \models \phi$, which means that $w_\Sigma \models C_\mathcal{B}\phi$.

The real work lies in the converse direction. Suppose that $w_\Sigma \models C_\mathcal{B}\phi$. The crucial idea is that we can find a sentence χ that has the properties that $\vdash \bigwedge\Sigma \to \chi$ and that $\vdash \chi \to \Box_a(\phi \wedge \chi)$ for each $a \in \mathcal{B}$. We can then use propositional logic and the rule RC to conclude that $\vdash \bigwedge\Sigma \to C_\mathcal{B}\phi$. By maximality and consistency of Σ, we then know that $C_\mathcal{B}\phi \in \Sigma$.

The sentence χ we are looking for is this one:

$$\bigvee\{\bigwedge\Gamma \mid w_\Gamma \models C_\mathcal{B}\phi\}$$

Since Φ is finite, each Γ is finite, and χ is indeed a sentence in our language. Clearly, $\bigwedge\Sigma \to \chi$. To see that $\chi \to \Box_a(\chi \wedge \phi)$, we first observe that for each disjunct $\bigwedge\Gamma$ of χ, it holds that $w_\Gamma \models C_\mathcal{B}\phi$, and therefore, $w_\Gamma \models \Box_a\phi$. It follows by standard reasoning that $\Box_a\phi \in \Gamma$. Since $\Box_a\phi$ is implied by each disjunct of χ, we conclude that $\vdash \chi \to \Box_a\phi$.

To show that $\chi \to \Box_a\chi$, assume to the contrary that $\chi \wedge \neg\Box_a\chi$ is consistent. By the fact that $\Box_a$ is a normal modal operator and that the sentence $\bigvee\{\bigwedge\Gamma \mid \Gamma$ is maximal consistent in $\Phi\}$ is a tautology, the sentence $\chi \wedge \neg\Box_a\chi$ is equivalent to $\bigvee\{\bigwedge\Delta \wedge \neg\Box_a\neg\bigwedge\Theta \mid w_\Delta \models C_\mathcal{B}\phi \text{ and } w_\Theta \not\models C_\mathcal{B}\phi\}$. If this sentence is consistent, there must be a Δ such that $w_\Delta \models C_\mathcal{B}\phi$ and a Θ such that $w_\Theta \not\models C_\mathcal{B}\phi$ with the property that $\bigwedge\Delta \wedge \neg\Box_a\bigwedge\Theta$ is consistent. But that means that $w_\Theta \in w_\Delta(a)$, which contradicts our assumption that $w_\Delta \models C_\mathcal{B}\phi$ but $w_\Theta \not\models C_\mathcal{B}\phi$.

Completeness now follows from the observation that if $\not\vdash \phi$, then there is a finite Fischer-Ladner closed set Φ that contains $\neg\phi$, and a set Σ that is maximal consistent in Φ such that $\phi \in \Sigma$. Then, $w_\Sigma \models \neg\phi$, so $\not\models \phi$.

The completeness of CK45 (and CS5) follows from the observation that the canonical models w_Σ are introspective (and factive) if we build the possibilities with CK45-consistent (CS5-consistent) sets. $\square$

To get an idea of what kind of sentences are valid in this logic, I'll present some theorems.

First of all, note that $C_\mathcal{B}$ is a normal modal operator: it satisfies the necessitation rule ($\vdash \phi$, then also $\vdash C_\mathcal{B}\phi$) and it distributes over implication (it holds that $\vdash C_\mathcal{B}(\phi \to \psi) \to (C_\mathcal{B}\phi \to C_\mathcal{B}\psi)$).

Consider now the limit case of common knowledge in a 'group' consisting of just one agent. In models where positive introspection holds, the differences between the common knowledge operator for singleton groups and the knowledge operators collapse: it holds that $\vdash_{\mathsf{CK45}} \Box_a\phi \leftrightarrow C_{\{a\}}\phi$. The reason is, of course, that if positive introspection holds, then $\Box_a\phi$ implies that $\Box_a^k\phi$ for each k, and by definition, that means that $C_{\{a\}}\phi$ is true.

Note also that if a sentence is common knowledge in a group $\mathcal{B}$ of agents, then it is also common knowledge in any smaller group: if $\mathcal{D} \subseteq \mathcal{B}$, then $\vdash_{\mathsf{CK}} C_\mathcal{B}\phi \to C_\mathcal{D}\phi$.

It holds that a sentence ϕ is common knowledge in $\mathcal{B}$ iff everyone in $\mathcal{B}$ knows that ϕ and knows that it is common knowledge that ϕ: the sentence $C_\mathcal{B}\phi \leftrightarrow \bigwedge_{a\in\mathcal{B}} \Box_a(\phi \wedge C_\mathcal{B}\phi)$ is a theorem in CK.

Another interesting fact is that 'positive introspection for common knowledge' $C_\mathcal{B}\phi \to C_\mathcal{B}C_\mathcal{B}\phi$ is true in all possibilities, regardless of the fact whether these possibilities are introspective or not. Common knowledge is always mutually accessible: if some fact is common knowledge, then everyone knows that it is common knowledge. In reflexive possibilities, the converse holds as well.

On the other hand, sentences of the form $\neg C_\mathcal{B}\phi \to C_\mathcal{B}\neg C_\mathcal{B}\phi$ are not valid in general, not even in introspective possibilities. The reason is that some sentence ϕ may fail to be common knowledge because one agent does not believe ϕ, while at the same time, another agent believes (falsely) that ϕ is common knowledge.

In belief models, agents are always fully informed about the common knowledge in the groups of which they are a member: $\vdash_{\mathsf{CK45}} \Box_a C_\mathcal{B}\phi \vee \Box_a\neg C_\mathcal{B}\phi$. In other words, in each possibility, and of each sentence ϕ, an agent either believes that ϕ is common knowledge, or he believes that ϕ is not common knowledge. He may be mistaken about it, but he never is in doubt about the question whether $C_\mathcal{B}\phi$ is the case.

In knowledge-models, it holds moreover that an agent cannot be mistaken. That means that $\vdash_{\mathsf{CS5}} C_\mathcal{B}\phi \leftrightarrow \Box_a C_\mathcal{B}\phi$ and $\vdash_{\mathsf{CS5}} \neg C_\mathcal{B}\phi \leftrightarrow \Box_a\neg C_\mathcal{B}\phi$.

3.3 Combining Information I: Distributed Knowledge

Consider a situation with two agents, a and b, and suppose that a has the information that p is the case, b knows that p implies q, but neither a nor b knows that q is the case. Even though neither one of the agents knows that q, there is a sense in which the information that q is the case is already present in their information states taken together: q is a logical consequence of the information that the two agents have. One way of formulating this is to say that the information that q is present in the 'system' consisting of both agents in a 'distributed' form: the information that q is distributed over information states of a and b. The standard term for such kind of knowledge is *distributed knowledge*.

The distributed knowledge between a and b is the information that a and b have 'together.' The topic of this section is to find a fair notion of 'adding' the information contained in one information state to the information contained in another one.

We can also put the question in a more concrete form. Suppose both a and b each have a certain amount of information, and communicate everything they know to a third agent, who initially has no information at all. What is the information state of this third agent c after a and b have communicated everything they know to c?

This question has received some attention in the literature. The term 'distributed knowledge' comes from, I believe, Halpern and Moses (1990). The notion is called 'implicit knowledge' in Halpern (1987). Humberstone (1985) gives the same truth definition for what he calls 'collective knowledge.' In Hilpinen (1969) and in Hilpinen (1974) the concept is used in deontic logic, where it is called the 'fiat-operator relative to $\Box$.'

Parts of this section appeared as Gerbrandy (1998).

Distributed Information as Intersection

A straightforward way of adding the information represented by two sets of possible worlds together is by taking the intersection of these two sets of worlds. In this conception, a sentence is distributed knowledge between a and b iff it is true in all worlds that occur in the information state of a as well as that of b. Fagin et al. (1995) put it as follows: "we combine the knowledge of the agents in group $\mathcal{B}$ by eliminating all worlds that some agent in $\mathcal{B}$ considers impossible."

Let us add a modal operator $D_{\mathcal{B}}$ to the language of classical modal logic, one for each $\mathcal{B} \subseteq \mathcal{A}$. The resulting language is denoted by $\mathcal{L}^D$. The intended interpretation of a sentence of the form $D_{\mathcal{B}}\phi$ is that ϕ is distributed knowledge among the agents in $\mathcal{B}$.

If (K, x) is a Kripke model and $\mathcal{B}$ is a set of agents, then the information distributed over the agents in $\mathcal{B}$ can be characterized by the set of worlds that are compatible with both the information of all agents in $\mathcal{B}$. In other words, the distributed knowledge in $\mathcal{B}$ is characterized by the set of worlds that are accessible from x both for each agent $a \in \mathcal{B}$ by the accessibility relation $\xrightarrow{a}$.

Definition 3.18 For Kripke models (K, x):[8]

$$(K, x) \models D_{\mathcal{B}}\phi \text{ iff } (K, y) \models \phi \text{ for all } y \text{ such that } x \xrightarrow{a} y \text{ for all } a \in \mathcal{B} \qquad \Box$$

It is perhaps not immediately obvious that the plausibility of this definition depends very much on the ontological view one has on possible worlds, and in particular what it means for two possible worlds to be different. To illustrate this point, consider a very simple model K with three worlds x, y and z, two agents a and b, an accessibility relation $\xrightarrow{a} = \{(x, y)\}$ and $\xrightarrow{b} = \{(x, z)\}$, and a valuation function that assigns to all propositional variables the same truth-values in y and z.

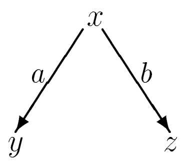


Figure 3.3.

In this model, the set of worlds that a considers possible is a singleton set containing one world y from which no further worlds are accessible. The set of worlds that b considers possible is also a singleton set containing one world, z, in which the same propositional variables are true as in y, and in which no further worlds are accessible.[9]

The information that a has in (K, x) is given by the singleton set $\{(K, y)\}$, and the information of b is given by the singleton set $\{(K, z)\}$. Since the two worlds are different, the intersection of the state of a with the state of b is empty. Therefore, $(K, x) \models D_{\{a,b\}}\bot$: the distributed knowledge of the two agents is inconsistent. On the other hand $\{(K, y)\}$ and $\{(K, z)\}$ have exactly the same

[8] Cf. Halpern and Moses (1990), Fagin et al. (1995)

[9] This model is not a belief- or knowledge-model. One can easily adapt the example to get a transitive and euclidean model by taking its transitive and euclidean closure: simply add a and b-edges $y \xrightarrow{a} y$, $y \xrightarrow{b} y$, $z \xrightarrow{b} z$ and $z \xrightarrow{a} z$ to the model. Changing the example into a proper model of knowledge, in which the accessibility relations are equivalence relations takes a little more work. Such an example can be found in Van der Hoek et al. (to appear). Since the point I want to make does not depend on these properties, I will use the more simple example.

structure, in the sense that their generated submodels are isomorphic.[10] The definition 3.18 of distributed knowledge, then, only makes sense if we take a view on Kripke models in which the difference between isomorphic worlds is somehow essential for the information represented by them: there must be something that distinguishes the world y from the world z that is relevant to the information of a and b in (K, x), and this distinction is not visible by looking at the value of the propositional variables and the information states of the agents alone.

Such a view on possible worlds is implied by the framework that is developed in Fagin et al. (1992) and in the fourth chapter of Fagin et al. (1995). I will discuss their model in detail in section 5.3. In their work, the indices in a Kripke model (the elements of W) have an internal structure themselves: they are descriptions of 'states', basically 'ways the world could be.' It is this internal structure of the possible world that makes the world 'what it is:' the values of the propositional variables and the accessibility relations in the Kripke model are an additional logical layer that makes it possible to speak about possible worlds in the language of epistemic logic, but this extra layer may represent the internal structure of the possible worlds only partially. If the structure of the Kripke model (that is, the model *modulo* the identity of the indices in that model) gives no clue as to how to distinguish two possible worlds, this is still no reason to consider them 'the same.' Under this conception, the distributed knowledge of a and b should indeed be inconsistent in the example of figure 3.3: after all, their information states contain *different* worlds.

The perspective on epistemic semantics that I have adopted in this dissertation is different. A possibility or a world in a Kripke model is a description of a possible way the world could be, and this description is completely exhausted by the values of atomic sentences and the information states assigned to each of the agents. With respect to possibilities, there can be no confusion about this: when two possibilities have the same structure, then they are the *same* possibility. Under such a conception, there is simply no relevant difference between the information state of a and of b in the example: the fact that their information states are modelled by different possible worlds is just a quirk of the model, not something that should influence the truth-value of sentences.

Consider now the following definition of distributed knowledge in possibilities:

Definition 3.19 For all possibilities w:

$$w \models D_{\mathcal{B}}\phi \quad \text{iff} \quad v \models \phi \text{ for each } v \in \textstyle\bigcap_{b \in \mathcal{B}} w(b) \qquad \square$$

The idea behind this definition is the same as that behind the definition of distributed knowledge in a Kripke model: distributed knowledge is characterized by

[10] Two models are isomorphic iff they are connected by a bisimulation that is one-one. If two models are isomorphic, they have exactly the same structure; only the identity of the elements in their respective domains may be different. If two models are isomorphic, they satisfy the same sentences of modal logic.

intersection. The results, however, are quite different. For example, let sol be the solution of the Kripke model of figure 3.3. We can picture the possibility $\mathsf{sol}(x)$ as in figure 3.4.

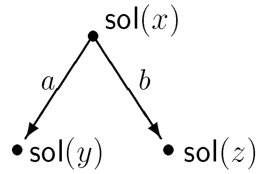


Figure 3.4.

Since z and y are bisimilar, they represent the same possibility, i.e. $\mathsf{sol}(z) = \mathsf{sol}(y)$. But that means that $\mathsf{sol}(x)(a) = \mathsf{sol}(x)(b) = \{\mathsf{sol}(y)\}$. In other words, the intersection of the information states of a and b in the solution of x is not empty, and therefore $\mathsf{sol}(x) \not\models D_{\{a,b\}}\bot$. The close correspondence between possibilities and their pictures —the preservation of the truth values of sentences— breaks down when we define the operator $D_{\mathcal{B}}$ using intersection.

The difference between the two definitions leads to a difference at the logical level as well. If our set of propositional variables is finite, we can describe the possibility that corresponds to the world x up to uniqueness by a single sentence. Let ψ_w be that sentence. Our example shows that $\neg D_{\{a,b\}}\bot$ is a semantical consequence of ψ_w with respect to all possibilities, but not with respect to all Kripke models. In other words, the choice of the framework we use to represent the information of agents influences the logical validities.

The matter is even further complicated by the fact that there are other views on epistemic semantics. Hintikka (1962) identifies possible worlds with maximal consistent sets of sentences. Transposing the definition of distributed knowledge as intersection in that framework leads to a logic, of distributed knowledge different from the other two. If one defines distributed knowledge in terms of 'intersection of accessible knowledge structures' (cf. chapter 2) or as 'intersection of situations' in situation theory, the result is yet another notion of distributed knowledge.

As an intermediate conclusion we can say that in the analysis of distributed knowledge in Kripke models as 'intersection of information states,' the truth value of sentences of the form $D\phi$ depends on the way the information states of the agents involved are modeled. The characterization is ontology-dependent, in the sense that different views on the role of possible worlds lead to different notions of distributed knowledge. I do not think this is a very big problem, but it does distinguish the definition of distributed knowledge from, say, the belief operator or the common knowledge operator, which are not in the same way ontology-dependent.

Adding Information as Logical Consequence

A way to avoid the problem of defining distributed knowledge semantically is to approach the question from a syntactic angle. We can combine two information states is by taking the logical consequence of sentences that are accepted in either state. With respect to distributed knowledge, this means that a sentence is distributed knowledge if and only if it is a logical consequence of the sentences that are known by the agents.

More formally, we can model this by taking all sentences of $\mathcal{L}$ that are accepted in either $w(a)$ or in $w(b)$, and say that ϕ is distributed knowledge between a and b just in case it is a logical consequence of this set of sentences.

Definition 3.20

$$w \models D_{\mathcal{B}}\phi \text{ iff } \{\psi \in \mathcal{L} \mid w \models \Box_a \psi \text{ for some } a \in \mathcal{B}\} \models \phi \qquad \Box$$

Humberstone (1985) characterizes what he calls 'collective knowledge' like this. The reason for looking at consequences of sentences of $\mathcal{L}$, as opposed to $\mathcal{L}^D$, is that in the latter case, the definition would be circular: the right hand side of the definition would quantify over all sentences believed by some agent in $\mathcal{B}$, which includes the sentence $D_{\mathcal{B}}\phi$ itself. Since $\mathcal{L}$ does not contain the operator D, we avoid this circularity.

We can also define distributed knowledge by logical consequence in Kripke models:

Definition 3.21

$$(K, x) \models D_{\mathcal{B}}\phi \text{ iff } \{\psi \in \mathcal{L} \mid (K, x) \models \Box_a \psi \text{ for some } a \in \mathcal{B}\} \models \phi \qquad \Box$$

These two definitions are 'equivalent', in the sense that possibilities and their pictures satisfy the same sentences of $\mathcal{L}^D$ under this definition:

Proposition 3.22 If (K, x) is a picture of w, then for each sentence of $\mathcal{L}^D$:

$(K, x) \models \phi$ under the satisfaction relation of definition 3.21 iff
$w \models \phi$ under the satisfaction relation of definition 3.20.

proof: By induction on the complexity of ϕ. $\Box$

This proposition suggests that the definition of distributed knowledge in terms of logical consequence does not depend on ontological positions in the same way as the definition in terms of intersection does. The effect of the definition depends on another parameter, however: that of the expressive power of the language. The stronger the language, the more sentences will be distributed knowledge. For example, suppose $w(a)$ and $w(b)$ are both singleton sets containing w_a and w_b respectively, and suppose that w_a and w_b are indistinguishable in the language $\mathcal{L}$ but distinguishable in a stronger language $\mathcal{L}^+$. In this case, relative to $\mathcal{L}$, the distributed knowledge of a and b is consistent (a and b consider the same sentences of $\mathcal{L}$ true), but relative to $\mathcal{L}^+$, their distributed knowledge is not consistent (there is a sentence of $\mathcal{L}^+$ of which a believes it is true, but b believes it is false).

If the notion of distributed knowledge between a and b is meant to capture the amount of information a and b can communicate, using the language $\mathcal{L}$, to a third agent, then the syntactic approach of definitions 3.20 and 3.21 seems to be the right one: this third agent will get exactly the information that can be expressed by the language in which the agents are communicating. The fact that the meaning of $D_\mathcal{B}$ depends on the expressive power of the language is very natural in this case: if the agents communicate using the language $\mathcal{L}$, then the amount of information that they can communicate depends on the expressive power of the language $\mathcal{L}$.

A Comparison of the Definitions

In this section I will compare the different definitions of distributed knowledge given above. The main conclusion is that although the different definitions give rise to different satisfaction relations for $\mathcal{L}^D$, the resulting consequence relations are all the same, under the assumption that the set of propositional variables in the language is infinite.

To simplify the discussion, I will use $\models_1$ for the interpretation of $\mathcal{L}^D$ as in definition 3.18, and I will use $\models_2$ to denote satisfaction relation defined in definition 3.20.

$$\begin{array}{lll}(K,x) \models_1 D_\mathcal{B}\phi & \text{iff} & \text{for all } y \text{ such that } x \xrightarrow{a} y \text{ and for each } a \in \mathcal{B} \\ & & \text{it holds that } (K,y) \models_1 \phi \\ (K,x) \models_2 D_\mathcal{B}\phi & \text{iff} & \{\psi \in \mathcal{L} \mid (K,x) \models \Box_a\psi \text{ for some } a \in \mathcal{B}\} \models \phi\end{array}$$

The comparison is divided in two subsections. Under the heading 'truth,' I will show that the two definitions do not assign the same truth-values to sentences, and identify two different classes of Kripke models in which the differences between the two definitions collapse. Under the heading 'logic,' there is a proof that if the set of propositional variables is infinite, then $\models_1$ and $\models_2$ give rise to the same logic, in the sense that the same sentences are valid under both conceptions.

Truth

If we consider the relations $\models_1$ and $\models_2$, it is not hard to see they are not the same. The following proposition shows that if we interpret the operator D by $\models_1$, it gets an interpretation that is weaker than in the interpretation under $\models_2$: any sentence of $\mathcal{L}$ that is distributed knowledge under the second conception is also distributed knowledge under the first:

Proposition 3.23 For all ϕ of $\mathcal{L}$:
If $(K,x) \models_2 D_\mathcal{B}\phi$ then $(K,x) \models_1 D_\mathcal{B}\phi$, but not vice versa.

proof: In this proof and the ones that follow, I will write x instead of (K, x) when this is not likely to lead to confusion, and write $x(a)$ for the set $\{(K, y) \mid x \xrightarrow{a} y\}$.

Assume $x \models_2 D_{\mathcal{B}}\phi$. Then $\{\psi \in \mathcal{L} \mid x \models \Box_a\psi \text{ for some } a \in \mathcal{B}\} \models \phi$. Take any $y \in \bigcap_{a\in\mathcal{B}} x(a)$. Clearly, for any ψ such that $x \models \Box_a\psi$ for some $a \in \mathcal{B}$, it holds that $y \models \psi$. So, by assumption, $y \models \phi$, and since y was arbitrary, it follows that $x \models_1 D_{\mathcal{B}}\phi$.

For the negative result, the model of our example on page 54 is an example of a model in which $(K, x) \models_1 D_{\{a,b\}}\bot$, but $(K, x) \not\models_2 D_{\{a,b\}}\bot$. □

In certain models, however, the operators are equivalent: in models that are *full* and in models that are *distinguishing*, the differences between the two operators collapse.

Proposition 3.24 (equivalence results)

1. A Kripke model K is *full* just in case for each x in K and each set of sentences Γ it holds that if $\{\psi \in \mathcal{L} \mid w \models \Box_a\psi \text{ for some } a \in \mathcal{B}\} \subseteq \Gamma$ and Γ is $\models_2$-satisfiable (i.e. there is a model in which all sentences of Γ are $\models_2$-true) then there is a $y \in \bigcap_{a\in\mathcal{B}} x(a)$ such that $y \models_2 \psi$ for all $\psi \in \Gamma$. If K is full, then:

$$(K, x) \models_1 \phi \text{ iff } (K, x) \models_2 \phi$$

   A similar result holds for possibilities.

2. A Kripke model K is *distinguishing* just in case for each x in K, each $y \in \bigcup_{a\in\mathcal{A}} x(a)$ and each $a \in \mathcal{A}$, there is a sentence ϕ_a of $\mathcal{L}$ such that $y \models \phi_a$ iff $y \in x(a)$.

   If K is distinguishing, then[11]

$$(K, x) \models_1 \phi \text{ iff } (K, x) \models_2 \phi$$

   A similar result holds for possibilities.

proof: For the first item, suppose that K is full. We prove the result by induction on ϕ, where the only interesting case is when ϕ is of the form $D_{\mathcal{B}}\psi$.

We show that for x in K it holds that:

$$(K, x) \models_1 D_{\mathcal{B}}\phi \text{ iff } (K, x) \models_2 D_{\mathcal{B}}\phi.$$

Assume that $x \not\models_2 D_{\mathcal{B}}\phi$. Then $\{\psi \in \mathcal{L} \mid x \models \Box_a\psi \text{ for some } a \in \mathcal{B}\} \not\models_2 \phi$. But then, the set $\Gamma := \{\psi \in \mathcal{L} \mid x \models \Box_a\psi \text{ for some } a \in \mathcal{B}\} \cup \{\neg\phi\}$ is $\models_2$-satisfiable, and since (K, x) is full this means that there must be some $x_\Gamma \in \bigcap_{a\in\mathcal{B}} x(a)$ in

[11] Van der Hoek et al. (to appear) prove a weaker result that gave me the idea for this one.

which all sentences of Γ are $\models_2$-true. But then in particular, $x_\Gamma \not\models_2 \phi$, so by induction hypothesis $x_\Gamma \not\models_1 \phi$, and therefore $x \not\models_1 D_{\mathcal{B}}\phi$.

For the other direction, the reasoning is the same as in proposition 3.23: suppose $x \models_2 D_{\mathcal{B}}\phi$. Then by definition, $\{\psi \in \mathcal{L} \mid x \models \Box_a\psi \text{ for some } a \in \mathcal{B}\} \models_2 \phi$. Then, *a forteriori*, for each $y \in \bigcap_{a\in\mathcal{B}} x(a)$, $y \models_2 \phi$, and therefore, by induction hypothesis, $y \models_1 \phi$.

For the second item, the proof is by induction on the number of occurrences of $D_{\mathcal{B}}$-operators in ϕ, with a subinduction on the structure of ϕ. The only interesting case in the induction is when ϕ is of the form $D_{\mathcal{B}}\phi$.

So, suppose K is distinguishing, and that we have proven the result for all sentences that contain at most as many occurrences of D as ϕ does.

Suppose first that $x \models_1 D_{\mathcal{B}}\phi$. Then, for each $y \in \bigcap_{a\in\mathcal{B}}(x)(a)$ it holds that $y \models_1 \phi$. Now take any $a \in \mathcal{B}$. Since K is distinguishing, there is a sentence ϕ_a of $\mathcal{L}$ such that for each u such that $x \xrightarrow{b} u$ for some b, it holds that $u \models \phi_a$ iff $x \xrightarrow{a} u$. So, clearly, $x \models_1 \Box_a\phi_a$, and, since $\phi_a \in \mathcal{L}$, also $x \models_2 \Box_a\phi_a$. Also, $x \models_1 \Box_a(\bigwedge_{b\in\mathcal{B}} \phi_b \to \phi)$. Since this sentence contains less occurrences of D than $D_{\mathcal{B}}\phi$ does, it follows that $x \models_2 \Box_a(\bigwedge_{b\in\mathcal{B}} \phi_b \to \phi)$.

Note that it holds for all χ and $a \in \mathcal{B}$ that if $x \models_2 \Box_a\chi$, then $x \models_2 D_{\mathcal{B}}\chi$. So, in particular, $x \models_3 D_{\mathcal{B}}\phi_a$ for each $a \in \mathcal{B}$, and $x \models_3 D_{\mathcal{B}}(\bigwedge_{a\in\mathcal{B}} \phi_a \to \phi)$. It also holds that if $x \models_2 D_{\mathcal{B}}\chi$ and $x \models_2 D_{\mathcal{B}}(\chi \to \chi')$, then $x \models_2 D_{\mathcal{B}}\chi'$. Combining all this, it follows that $x \models_2 D_{\mathcal{B}}\phi$, as we wanted to prove.

The other direction goes as before. $\Box$

The condition of being 'full' is subsumed under the property of Kripke models that information states can be characterized by a set of sentences, i.e. that information states consist of *all* models of a particular set of sentences. If we assume that the beliefs of the agent can be expressed in the object language, then this is a natural consequence of the slogan that 'the beliefs of an agent are modeled by the set of possibilities compatible with his beliefs.' The condition of being 'distinguishing' is subsumed under the property that information states consist of all models of some *finite* set of sentences.

The results above show that the differences between $\models_1$ and $\models_2$ are relevant only in models in which information states cannot be characterized by a set of sentences.

Logic

We have seen above that when our set of propositional variables is finite, the logic of $\models_1$ is not the same as the logic of $\models_2$. When we have infinitely many propositional variables, however, the logics of both relations are the same, in the sense that any sentence that is valid under $\models_2$ is also valid under $\models_1$, and vice versa. We can show this by using the properties of fullness and being distinguishing of the previous section. The results also apply to validity with respect to possibilities, for either definition.

Proposition 3.25 If $\mathcal{P}$ is infinite, then the following four statements are equivalent for all sentences ϕ of $\mathcal{L}^D$.

1. For all (K, x): $(K, x) \models_1 \phi$
2. For all (K, x): $(K, x) \models_2 \phi$
3. For all w: $w \models_1 \phi$
4. For all w: $w \models_2 \phi$

Moreover, this equivalence holds also if we restrict the quantification to all introspective and/or reflexive models.

proof:
[1 $\Rightarrow$ 2] Suppose $(K, x) \not\models_2 \phi$. Then, with lemma 3.26, we can find a full model (K', x') such that $(K', x') \not\models_2 \phi$. But then, $(K', x') \not\models_1 \phi$.

[2 $\Rightarrow$ 1]. Suppose there is a K such that $(K, x) \not\models_1 \phi$. Then we can use lemma 3.27 and find a distinguishing model (K', x') such that $(K', x') \not\models_1 \phi$. But if (K', x') is distinguishing, this implies that $(K', x') \not\models_2 \phi$.

[1 $\Rightarrow$ 3] Suppose that there is a w such that $w \not\models_1 \phi$. There is an extensional picture (K, x) of w for which it holds with lemma 3.29 that $(K, x) \not\models_1 \phi$.

[3 $\Rightarrow$ 1] Suppose $(K, x) \not\models_1 \phi$. Let, for each possible world y in K, p_y be a distinguished propositional variable that does not occur in ϕ. We can construct a new model K' that differs from K only in that $V(p_y) = \{y\}$ for each y in the domain of K. Since none of the variables p_y occur in ϕ, it holds that also $(K', x) \not\models \phi$. Now, since for each world in the domain of K' there is a propositional variable true at that world only, it is easy to see that the solution sol of (K', x) is such that (K', x) is an extensional picture of $\mathsf{sol}(x)$. Using lemma 3.29 again, we conclude that $\mathsf{sol}(x) \not\models_1 \phi$.

[2 $\Leftrightarrow$ 4] This follows immediately from proposition 3.22. $\square$

In the proof, I used the following two lemmas.

Lemma 3.26 For each (K, x) there is a full model $(K, x)'$ such that $(K, x) \models_2 \phi$ iff $(K, x)' \models_2 \phi$. Moreover, we can find a $(K, x)'$ that is euclidean, transitive or reflexive just in case (K, x) is.

proof: The proof is just a simple variation on the canonical model construction of the standard completeness proof (cf. proposition 3.30). Define K' as follows. For its domain, K' has all maximal $\models_2$-satisfiable sets, i.e. all sets of sentences Σ that are satisfiable and are maximal in the sense that adding any sentence to Σ results in a set that is not satisfiable. The accessibility relations of K' are given by: $\Sigma \xrightarrow{a} \Gamma$ iff for all ψ, if $\Box_a \psi \in \Sigma$, then $\psi \in \Gamma$, and the valuation function assigns to each Σ and p the value 1 just in case $p \in \Sigma$.

For a euclidean, transitive or reflexive model, we construct K' with sets that are satisfiable in the appropriate models.

It holds that:

$$(K', \Sigma) \models_2 \phi \text{ iff } \phi \in \Sigma$$

The proof works by first showing the result for all sentences of $\mathcal{L}$ by induction on ϕ, and then showing that the result holds for all sentences of $\mathcal{L}^D$, again by induction on ϕ. The details of the proof are similar to those in the standard Henkin proof of the completeness of classical modal logic. The case where ϕ is of the form $D_{\mathcal{B}}\psi$ runs as follows:

$(K', \Sigma) \models_2 D_{\mathcal{B}}\psi$ iff
$\{\chi \in \mathcal{L} \mid (K', \Sigma) \models \Box_a\chi \text{ for some } a \in \mathcal{B}\} \models_2 \psi$ iff (induction hypothesis)
$\{\chi \in \mathcal{L} \mid \Box_a\chi \in \Sigma \text{ for some } a \in \mathcal{B}\} \models_2 \psi$ iff (since Σ is maximal)
$D_{\mathcal{B}}\psi \in \Sigma$

Clearly, the set of sentences that are true in (K, x) is satisfiable. Let Σ be this set. Then (K', Σ) is a full model such that $(K, x) \models_2 \phi$ iff $(K', \Sigma) \models_2 \phi$ for each $\phi \in \mathcal{L}^D$. $\Box$

Lemma 3.27 For each (K, x) such that $(K, x) \models_1 \phi$, there is a distinguishing model (K', x') such that $(K', x') \models_1 \phi$.

Moreover, if K is transitive and euclidean (and reflexive), then we can choose K' to be transitive and euclidean (and reflexive) as well.

proof: Suppose $(K, x) \models_1 \phi$. We know from the completeness proof that we can find a countable model (K', x') such that $(K', x') \models_1 \phi$. Let, for each y in the domain of K', $p_{a,y}$ be a propositional variable that does not occur in ϕ. Here we use the fact that the set of propositional variables is infinite.

Now let K'' be exactly as K, except that $V(z)(p_{y,a}) = 1$ iff $y \xrightarrow{a} z$. This makes K'' distinguishing by definition. Of course, if K' is transitive, euclidean or reflexive, then so is K''.

Since ϕ does not contain any of the propositional variables $p_{a,y}$, it follows by a standard argument that $(K', x) \models \phi$ iff $(K'', x) \models \phi$. So, (K'', x) is the model we are looking for. $\Box$

There is a certain class of Kripke models that correspond with possibilities in such a way that truth of sentences of $\mathcal{L}^D$ *is* preserved. If a Kripke model is an *extensional* picture of a possibility, then the same sentences will be true in both models:

Definition 3.28 (extensional pictures)
A Kripke model (K, x) is an *extensional picture* of a possibility w iff (K, x) is a picture of w with a solution that is one-one. $\Box$

Lemma 3.29 If (K, x) is an extensional picture of w, then for all sentences ϕ of $\mathcal{L}^D$:

$$(K, x) \models \phi \text{ under the satisfaction relation of definition 3.18 iff}$$
$$w \models \phi \text{ under the satisfaction relation of definition 3.19}$$

proof: By induction on ϕ, where the only interesting case is when ϕ is of the form $D_{\mathcal{B}}\phi$. The crucial observation here is that if (K, x) is an extensional picture of w, then it holds that there is a $v \in \bigcap_{a \in \mathcal{B}} w(a)$, if and only if there is a y such that (K, y) is an extensional picture of v and $x \xrightarrow{a} y$ for each $a \in \mathcal{B}$. □

We have proven that $\models_1$ and $\models_2$ give rise to the same notion of validity. I state here, without proof, that this result can be extended: the set of sentences of $\mathcal{L}^D$ true in all Hintikka models, or in all possibilities, is the same set as the set of $\models_1$ (or $\models_2$) valid sentences.

Completeness

We have seen that, if our language has infinitely many propositional variables, the two notions of validity $\models_1$ and $\models_2$ are the same. In the following, we will write $\models_{\mathsf{DK}} \phi$ iff $(K, x) \models_1 \phi$ for each (K, x), and $\models_{\mathsf{DK45}} \phi$ and $\models_{\mathsf{DS5}} \phi$ for, respectively, their transitive and euclidean, and transitive, euclidean and reflexive counterparts.

Adding the following two axioms to the logic of K provides a sound and complete axiomatization of all valid sentences of $\mathcal{L}^D$.

D1 $D_{\{a\}}\phi \leftrightarrow \Box_a\phi$.

D2 $D_{\mathcal{B}}(\phi \to \psi) \to (D_{\mathcal{C}}\phi \to D_{\mathcal{D}}\psi)$ if $\mathcal{B} \subseteq \mathcal{D}$ and $\mathcal{C} \subseteq \mathcal{D}$.

We let DK consist of the axioms of K together with D1 and D2.

Axiom D1 says that the knowledge that is distributed in the 'group' consisting of a only is just the knowledge of a. Axiom D2 says that if a certain group has distributed knowledge of ϕ, and another group has distributed knowledge that ϕ implies ψ, then both groups together have distributed knowledge of ψ as well.

To consider one implication of these axioms: if $\mathcal{B} \subseteq \mathcal{C}$, then $\vdash D_{\mathcal{B}}\psi \to D_{\mathcal{C}}\psi$ for all ψ.[12] This validity corresponds with the intuition that if a sentence is distributed knowledge in a group $\mathcal{B}$, it will also be distributed knowledge in any group larger then $\mathcal{B}$. One can see this as a generalization of the maxim that 'two know more than one.'

[12] If $\mathcal{B} \subseteq \mathcal{C}$, and $a \in \mathcal{B}$, then $\vdash D_{\{a\}}(\psi \to \psi) \to (D_{\mathcal{B}}\psi \to D_{\mathcal{C}}\psi)$ is a special case of D2. With axiom D1, it follows that $D_{\{a\}}(\psi \to \psi)$, so, by propositional logic, we can conclude that $\vdash D_{\mathcal{B}}\psi \to D_{\mathcal{C}}\psi$

Also, these axioms imply that $D_{\mathcal{B}}$ is a normal modal operator, in the sense that a necessitation rule for $D_{\mathcal{B}}$ is a derived rule[13] and that D distributes over implication.[14]

The logic DK45 is given by adding the following two axioms to DK:

D4 $D_{\mathcal{B}}\phi \to D_{\mathcal{B}}D_{\mathcal{B}}\phi$.

D5 $\neg D_{\mathcal{B}}\phi \to D_{\mathcal{B}}\neg D_{\mathcal{B}}\phi$.

The logic DS5 is given by adding the following axiom to DK45:

DT $D_{\mathcal{B}}\phi \to \phi$.

Proposition 3.30 (Completeness)
DK is a sound and complete axiomatization of $\models_{\mathsf{DK}}$.
DK45 is a sound and complete axiomatization of $\models_{\mathsf{DK45}}$
DS5 is a sound and complete axiomatization of $\models_{\mathsf{DS5}}$.

proof: I'll present a sketch of the proof. It is a generalisation of the completeness proof of DS5 from Fagin et al. (1992).

Let a 'pseudomodel' be a model of the form $(W, (\xrightarrow{\mathcal{B}})_{\mathcal{B}\subseteq\mathcal{A}}, V)$, and define 'pseudo-satisfiability' as a relation between sentences of $\mathcal{L}^D$ and pseudo-models by treating the operators $D_{\mathcal{B}}$ as quantifying over $\xrightarrow{\mathcal{B}}$-accessible worlds. With respect to pseudo-satisfiability, the operators $D_{\mathcal{B}}$ are just classical modal operators. We can use standard techniques from the completeness proof of classical modal logic to show that any DK (or DK45 or S5) consistent theory can be pseudo-satisfied in a pseudo-model (which is transitive, euclidean and reflexive, if necessary). It is also not very hard to check that the rules are sound for pseudo-models with the property that if $\mathcal{B} \subseteq \mathcal{C}$, then $\xrightarrow{\mathcal{B}}\subseteq\xrightarrow{\mathcal{C}}$. The canonical model has this property. Therefore, the logic is sound and complete with respect to pseudo-models with the property that if $\mathcal{B} \subseteq \mathcal{C}$, then $\xrightarrow{\mathcal{B}}\subseteq\xrightarrow{\mathcal{C}}$. Since $D_{\mathcal{B}}$ is just another classical modal operator in these models, it follows that we can construct finite canonical models, and so the logic is decidable.

We can also use standard techniques to show that any pseudo-model can be unraveled into a model that looks like a tree, and in which the same sentences are true.

Now turn the unraveled pseudo-model into a model for $\mathcal{L}^D$ by taking the same worlds and valuation function, and setting $w \xrightarrow{a} v$ in the new model just in case there is a $\mathcal{B}$ such that $a \in \mathcal{B}$ and $w \xrightarrow{\mathcal{B}} v$ in the pseudo-model. We can now show that any sentence that is pseudo-satisfied at some world x in the restored unraveled pseudo-model is satisfied at x in the new model.

[13] Assume that $a \in \mathcal{B}$. Then $\vdash \psi$ implies that $\vdash \Box_a\psi$, which implies by D1 that $\vdash D_{\{a\}}\psi$, which implies by D2 that $\vdash D_{\mathcal{B}}\psi$.

[14] This is a special case of D2, where the sets $\mathcal{B}$, $\mathcal{C}$ and $\mathcal{D}$ are all the same.

We have now proven that DK is complete with respect to all models. Since we need only an unraveled pseudo-model of finite 'depth' to satisfy a given sentence (the length of the longest path in the tree need not be greater than the maximal nesting of modal operators in the sentence), DK has the finite model property.

The new model does not have the properties associated with belief or knowledge. To show that every DK45-consistent set is satisfiable in a belief model, we can simply take the transitive and euclidean closure of the accessibility relations in the tree model, and for DS5, we take the reflexive transitive and euclidean closure. Lemma 3.31 guarantees that this can be done: the new model is safe. □

Lemma 3.31 Let $K = (W, (\stackrel{a}{\longrightarrow})_{a \in \mathcal{B}})$ be a model such that for each $x \in W$, the set of sentences that are true at x is DK45-consistent. We define the relation $\stackrel{\mathcal{B}}{\leadsto}$, for each $\mathcal{B} \subseteq \mathcal{A}$ relative to K, as follows:

$y \stackrel{\mathcal{B}}{\leadsto} z$ iff there are $y_0 \ldots y_n$ and $y'_1 \ldots y'_n$ such that for each $a \in \mathcal{B}$: $y_0 \stackrel{a}{\longrightarrow} y_1 \stackrel{a}{\longrightarrow} \ldots \stackrel{a}{\longrightarrow} y_n$, $y_0 \stackrel{a}{\longrightarrow} y'_1 \stackrel{a}{\longrightarrow} \ldots \stackrel{a}{\longrightarrow} y'_m$ (with $n \geq 0$ and $m \geq 1$) and $y_n = y$ and $y'_m = z$. Note that $\stackrel{a}{\leadsto}$ is the transitive and euclidean closure of $\stackrel{a}{\longrightarrow}$.

We say that K is *safe* when $y \stackrel{\mathcal{B}}{\leadsto} z$ iff for each $a \in \mathcal{B}$, $y \stackrel{a}{\leadsto} z$. Models that look like a tree are safe, for example, and also models that are transitive and euclidean.

In safe models it holds that:

$$(W, (\stackrel{a}{\leadsto})_{a \in \mathcal{B}}, V), x) \models \phi \text{ iff } (W, (\stackrel{a}{\longrightarrow})_{a \in \mathcal{B}}, V), x) \models \phi.$$

proof: By induction on ϕ. The interesting case is when ϕ is of the form $\Box_a \psi$, where we use the fact that the theory of x is DK45-consistent. Once we have this, the case where ϕ is of the form $D_{\mathcal{B}}\psi$ is straightforward with the assumption that K is safe. □

The literature contains a whole range of proofs for modal logics 'with intersection.' The following is an attempt at an overview. There is a very elegant completeness proof in Gargov and Passy (1990) for a logic they call 'Boolean modal logic,' of which the language $\mathcal{L}^D$ is only a small fragment. They also show that DK has the finite model property. Their proof works just as well for our case. Passy and Tinchev (1991) also study a richer language than the one considered here. They give a completeness proof that may also work for DK. Since the two last-mentioned articles get their inspiration from propositional dynamic logic as opposed to epistemic logic, these proofs apply only to the logic DK; the authors are not concerned with proving completeness with respect to transitive or euclidean models, that are typical of epistemic semantics. It turns out that these proofs are much more difficult. Fagin et al. (1992) give a proof of the completeness of DS5; Fagin et al. (1995) claim that the logics DK and DK45 are complete as well, for the language with all $D_{\mathcal{B}}$-operators for each $\mathcal{B} \subseteq \mathcal{A}$, but they do not give a detailed proof. Van der Hoek and Meyer (1992) and Van der Hoek and Meyer

(1997) contain completeness proofs for DK45 and DK as well, but only for a language with a single operator $D_{\mathcal{A}}$, with $\mathcal{A}$ the set of *all* agents. Since their proofs are rather long and opaque, it is not immediately obvious how their techniques can be used to work for the full language $\mathcal{L}^D$. Finally, Yde Venema (personal communication) has a proof of the completeness of DK using the 'step-by-step' method, that can be extended to cover the cases of DK45 and DS5 as well.

Conclusions

I have compared two different schemes for defining the semantics of an operator that expresses distributed knowledge in a Kripke model. In the first scheme, the operator is defined as a quantifier over the intersection of the accessibility relations of the agents involved. This definition was argued to be dependent on the identity criteria between possible worlds. The second definition of distributed knowledge in terms of logical consequence turned out to be dependent on the expressive power of the language. However, under certain conditions the differences between the definitions disappear, in particular, when information states always consist of all models of a certain set of sentences. Moreover, if the language contains an infinite amount of atomic variables, the different semantics have the same weakly sound and complete axiomatization.

These results show that the issue of what a good definition of distributed knowledge is cannot be decided on the basis of logic alone. We cannot choose the one definition over and above another on the basis of our intuitions about which sentences logically follow from others, because in this respect, the definitions are equivalent. This means that the issue must be decided on the level of the semantics.

The main focus in this dissertation will be on possibilities as a model for epistemic logic. Therefore, in the remainder of this dissertation the definition of distributed knowledge as intersection is taken as basic. I repeat it here:

Definition 3.32 (distributed knowledge)
Let w be a possibility. Then *the distributed knowledge in $\mathcal{B}$ in w*, $D_{\mathcal{B}}(w)$, is the set $\bigcap_{b\in\mathcal{B}} w(b)$, and we define satisfaction in possibilities by:

$$w \models D_{\mathcal{B}}\phi \quad \text{iff} \quad \text{for all } v \in D_{\mathcal{B}}(w)\text{: } v \models \phi \qquad \square$$

3.4 Combining Information II: Combined Knowledge

One way to think of the distributed knowledge in a group of agents is by seeing it as the information that a third agent would have if he had the information of each of the agents. The distributed information of a group of agents is all that information they can, ideally, communicate to an outside observer. A closely related notion is that of the information each of the agents would have if they combined their information together by communicating to each other.

This notion is different from that of distributed knowledge. Suppose an agent a knows that p is the case, and knows also that another agent b does not know wether p. In such a situation, the statement that b does not know that p is distributed knowledge. Suppose now that a tells b that p is the case. If all goes well, the result will be that b knows that p is the case. Moreover, if b's information is introspective, he will also know that he knows that p. Either a might infer that b knows that p, or b might tell a himself that he now knows that p. Eventually, both a and b will know that b knows that p. So, the fact that b knows that p will be known by both of the agents after communicating their information. But the statement that b knows that p is *not* distributed knowledge in the initial situation.

I will use the term 'combined knowledge' (of a certain group of agents $\mathcal{B}$) for the information that results from the agents (in $\mathcal{B}$) sharing their knowledge together. As the example suggests, it is the notion of combined knowledge, and not so much that of distributed knowledge, that is of interest in a theory of communication.[15] The combined knowledge in a certain situation is the information that results after the agents have communicated all their information to each other. This means that if a certain fact is combined knowledge, then it should remain combined knowledge after information is 'properly exchanged.' Vice versa, no new facts will become combined knowledge as the result of a proper exchange of information. I will study these ideas in more detail in chapter 6.5.

Let us consider the matter in more detail, and fix two agents, a and b and a possibility w. What information would a and b have after they manage to communicate all the information they have in w?

First of all, this would be a situation in which a and b have the *same* information: if not, there would still be information one of the agents has, but not the other.

Secondly, each agent would know exactly which information each of the other agents has (if we assume that the agents have introspective information). If not, then there would still be information to communicate. So, after full communication took place, both a and b will be in an information state σ, in which it is

[15]I want to stress this point, because the distinction is not always noted. For example, Hoek et al. (to appear) write "From the point of view of communicating agents, G-knowledge [which is here called distributed knowledge] may be seen as the knowledge being obtained if the agents were fully able to communicate with each other. Actually, instead of being able to communicate with each other, one may also adopt the idea that the G-knowledge is just the knowledge of one distinct agent, to whom all the agents communicate their knowledge." In Borghuis (1994) one finds: "Implicit knowledge [i.e. what is here called distributed knowledge] is of interest in connection with information dialogues: if we think of the dialogue participants as agents with information states represented by epistemic formulae, then implicit knowledge precisely defines the propositions the participants could conclude to during an information dialogue."

known that both a and b have the information represented by σ. Formally, it holds that for each $w \in \sigma$, $w(a) = w(b) = \sigma$.[16]

What are the possibilities in the 'combined state' σ? Consider the following scenario. Suppose the agents a and b have a roundabout way of communicating information to each other: they first give all the information they have to another, 'neutral,' agent, who collects all this information together, and then makes a 'public broadcast' to the agents a and b of this information. The effect of such a public broadcast is that both a and b obtain the information broadcasted, as well as the information that both a and b know that there was such a public broadcast.[17]

The information that a and b give to the neutral agent is of course exactly the information that is distributed between a and b. This information is represented by $D_{\{a,b\}}(w)$ (in whatever way we define it; see the previous section). The overall effect of a public broadcast of the information distributed between a and b is that both agents exclude all possibilities not in $D_{\{a,b\}}(w)$ from their original information state, and then adapt each of the remaining possibilities to the effect that both a and b have received the same information. If a is in state $w(a)$, the resulting state σ of a will contain all possibilities in $w(a) \cap D_{\{a,b\}}(w)$, except that each of these possibilities should be changed in such a way that both a and b are in the state resulting after the public broadcast. The same holds for b.

If this is right, then a's new state σ_a will have the following property:[18]

$$\sigma_a = \{w \mid \exists v \in w(a) \cap D_{\{a,b\}}(w) : v[a,b]w \text{ and } w(a) = \sigma_a \text{ and } w_b = \sigma_b\}$$

Similarly, b's state σ_b satisfies the condition that:

$$\sigma_b = \{w \mid \exists v \in w(b) \cap D_{\{a,b\}}(w) : v[a,b]w \text{ and } w(a) = \sigma_a \text{ and } w(b) = \sigma_b\}$$

Note that $w(a) \cap D_{\{a,b\}}(w) = D_{\{a,b\}}(w) = w(b) \cap D_{\{a,b\}}(w)$. Given that σ_a and σ_b satisfy the two equations above, it is now not very hard to see that $\sigma_a = \sigma_b = \{w \mid \exists v \in D_{\{a,b\}}(w) : v[a,b]w \text{ and } w(a) = w(b) = \sigma_a\}$.

Generalizing, we can say that for any group of agents $\mathcal{B}$, the information state σ of each of the agents after the public broadcast of the information in $D_{\mathcal{B}}(w)$ is a state σ that consists of all possibilities in $D_{\mathcal{B}}(w)$, changed in such a way that each of the agents in $\mathcal{B}$ also has the information that $D_{\mathcal{B}}(w)$.

Definition 3.33 The *combined information* of $\mathcal{B}$ in w, denoted by $I_{\mathcal{B}}(w)$, is the unique state σ for which it holds that:

[16] This means that σ is a common ground in the sense of Zeevat (1997). Cf chapter 6.

[17] Using the terminology from chapter 4, a and b perform a mutual update with the information broadcasted.

[18] I write $w[a,b]v$ iff w and v differ at most in the information state they assign to a and b. Similarly, $w[\mathcal{B}]v$ means that w and v differ at most in the information states they assign to agents in $\mathcal{B}$.

$$\sigma = \{v \mid \exists u \in D_{\mathcal{B}}(w) : u[\mathcal{B}]v \text{ and } v(b) = \sigma \text{ for each } b \in \mathcal{B}\}$$

We can now add operators $I_{\mathcal{B}}$ to the language with the following semantics:

$$w \models I_{\mathcal{B}}\phi \quad \text{iff} \quad \forall v \in I_{\mathcal{B}}(w) : v \models \phi$$ □

In chapter 6.5, I will discuss a simple dialogue game. We will see there that under certain assumptions on communication (most importantly, that all agents only communicate things they believe) the notion of combined information behaves as we would like it to behave. In such dialogues the states of each of the agents involved will converge to the state that represents their combined information. Under certain assumptions (namely that all the information that is represented by their information states can in fact be communicated – if their states are *full* in the sense of definition 3.24), they will actually reach the state of combined information. At the same time, and also this seems right, it holds that the combined information of a certain set of agents stays the same during the communication process.

Combined and Distributed Knowledge

The operator $I_{\mathcal{B}}$ behaves like a normal modal operator: its semantics satisfies the necessitation rule and the operator distributes over implication:

NecI If $\vdash \phi$, then $\vdash I_{\mathcal{B}}\phi$

I0 $\vdash I_{\mathcal{B}}(\phi \to \psi) \to (I_{\mathcal{B}}\phi \to I_{\mathcal{B}}\psi)$

In addition to these axioms, the following two axioms are valid.

I1 $\vdash D_{\mathcal{B}}\phi \leftrightarrow I_{\mathcal{B}}\phi$, if ϕ does not contain any operators with agents in $\mathcal{B}$ in their subscript.

I2 $\vdash I_{\mathcal{B}}\Box_a\phi \leftrightarrow I_{\mathcal{B}}D_{\mathcal{C}}\phi \leftrightarrow I_{\mathcal{B}}I_{\mathcal{C}}\phi \leftrightarrow I_{\mathcal{B}}C_{\mathcal{C}}\phi \leftrightarrow I_{\mathcal{B}}\phi$, if $a \in \mathcal{B}$ and $\mathcal{C} \subseteq \mathcal{B}$

The first axiom states that for sentences that are not about the information of the agents in $\mathcal{B}$, combined information reduces to distributed knowledge. In other words, the differences between the two operators are only relevant when we consider higher-order information about the agents in $\mathcal{B}$ themselves.

The second axiom states that in the scope of an operator $I_{\mathcal{B}}$, all distinctions between truth, knowledge, common knowledge, distributed knowledge and combined knowledge of agents in $\mathcal{B}$ collapse. After full communication took place, all the information that each of the agents is public information (all knowledge is common knowledge), and there is no more knowledge to share between the agents (so everything that is combined knowledge or distributed knowledge is known by each of the agents). Among other things, this implies that the combined knowledge of the group $\mathcal{A}$ of all agents is completely determined by the propositional sentences that are combined knowledge.

These axioms together with the axioms for the operators $D_{\mathcal{B}}$ provide a sound and complete axiomatization of the sentences in which $I_{\mathcal{B}}$ only occurs if $\mathcal{B}$ is the set of all agents.

Proposition 3.34 (completeness for $I_{\mathcal{A}}$)
Let $\mathcal{L}^{CDI}$ be the language of epistemic logic with operators $C_{\mathcal{B}}$, $D_{\mathcal{B}}$ for each $\mathcal{B} \subseteq \mathcal{A}$, and an operator $I_{\mathcal{A}}$, where $\mathcal{A}$ is the set of all agents.

The axioms I0, I1 and I2 together with the necessitation rule and the axioms of DCK (DCK45, DS5 respectively) provide a sound and complete axiomatization of the sentences true in all possibilities (that are introspective and euclidean, and introspective, euclidean and factive, respectively).

proof: Completeness follows easily with the completeness of DCK. With axiom I2 and the fact that $I_{\mathcal{A}}$ is a normal modal operator, it is straightforward to define a function that translates all sentences of the form $I_{\mathcal{A}}\phi$ into equivalent sentences of the from $I_{\mathcal{A}}\psi$, where ψ does not contain any epistemic operators at all. With axiom I1, we know that the sentence obtained in this way is equivalent to $D_{\mathcal{A}}\psi$, and we can use the completeness result of DCK to conclude that the new logic is complete. □

So far, I have not been able to find a complete axiomatization for all validities of sentences of the full language that includes all operators of the form $I_{\mathcal{B}}$. However, I would like to make a few small observations about the logic. The axiom I2 already stated that under the scope of the operators $I_{\mathcal{B}}$, the distinctions between group knowledge of agents in $\mathcal{B}$, and the knowledge of separate agents collapses completely. The following axiom corroborates this interpretation of the axiom:

$\vdash I_{\mathcal{B}} I_{\mathcal{C}} \psi \leftrightarrow I_{\mathcal{B}} I_{\mathcal{C} \cup \mathcal{B}} \psi$, if $\mathcal{B} \cap \mathcal{C} \neq \emptyset$.

Finally, I would like to mention that if we consider just a single agent with introspective information, the differences between combined information and simple belief collapses: a single agent has nothing new to communicate to herself.

$\vdash_{\mathsf{belief}} \Box_a \psi \leftrightarrow I_{\{a\}} \psi$.

We obtained similar results for common knowledge (in K45-models) and distributed knowledge: all the group operators reduce to the belief operator when we consider only a single agent.

3.5 Relativised Knowledge

In most actual situations in which human agents are involved, there will often be at least one sentence about which two agents disagree. This is a problem if we want to use the definitions of distributed and combined knowledge for modeling such situations. If there is just one sentence about which two agents disagree, the information states modeling their beliefs will be completely disjoint. This means that their distributed and combined information is inconsistent: the notion trivializes and becomes useless. Strictly speaking, this is correct: if the agents put all their information together and there is one sentence about which they disagree, the result is an inconsistent theory.

Still, in real life, people are able to put their differences aside (if temporarily) and concentrate on the matters at hand. If your goal is to fix your bike together with a friend, it is irrelevant that the information distributed between you is inconsistent because he happens to believe that the moon is not made of cheese, while you know that it is. What is relevant is the knowledge you and your friend have about the topic of bike-fixing, either separately or distributed between you. Your convictions about astronomy are irrelevant in such a case. The point is that when agents communicate, they usually do not share all their knowledge, but only a certain part of it that is relevant to the situation at hand. A definition of combined knowledge that incorporates this is more useful than one that does not.

In this section, I want to propose another, less absolute way of approaching the concept, and sketch one way of making formal sense of the notion of information about a certain topic. To model the topic of a dialogue, I will use the semantics for questions of Groenendijk and Stokhof (1984) and Groenendijk and Stokhof (1997). In this model, the meaning of a question is equated with the set of its 'complete answers.' If one assumes that for each possibility, a question has one unique complete true answer, and assumes moreover that the meaning of an answer is a proposition (i.e. a set of possible worlds), this view naturally leads to the representation of the meaning of a question as a 'partition of the logical space.' A partition divides up the set of possible worlds into a set of pairwise disjoint classes, each of which represent a complete answer to the question.

Definition 3.35 (partitions)
A partition Π is a set of sets of possibilities such that $\bigcup \Pi$ is the class of all possibilities, $\emptyset \notin \Pi$, and if $\pi, \pi' \in \Pi$, then $\pi \cap \pi' = \emptyset$. □

For reasons that will become clear, I will refer to the elements of a partition as the *blocks* of that partition. A simple yes-no question such as "Is it raining now?" has two complete answers: "It is raining" and "It is not raining." The corresponding partition consists of the two propositions expressed by these sentences. For another example, suppose there are only two different objects that possibly might go to the party tonight: John and Mary, and consider the question "Who is coming to the party tonight?" In the restricted domain of would-be party-goers that we consider here, this question has four possible complete answers: "No-one," "John and no-one else," "Mary and no-one else," and "John and Mary." As a final example, consider the question "How many stars are there in heaven?" The possible complete answers to this question are "None," "Exactly one," "Exactly two," etcetera.

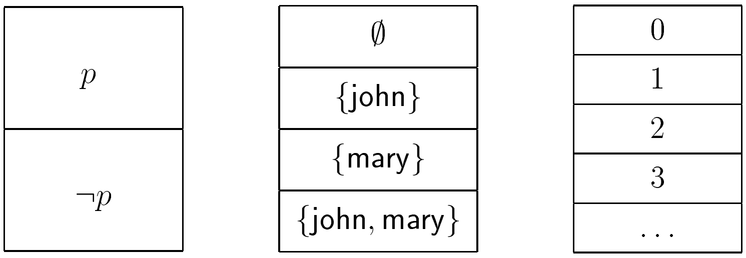


Figure 3.5.

One of the nice properties of this analysis of questions is that partitions have a friendly graphical representation. The three examples given above can be pictured as in figure 3.5. In these pictures, the outer box represents the logical space, and each box inside it represents a complete answer to the question. The fact that two possibilities are elements of the same block in the partition means that the differences between these two possibilities are irrelevant to the question at hand.

This semantics of questions can be combined with epistemic semantics in a very natural way. For any information state σ, we can determine how much information σ contains about a partition Π by checking how many answers in Π are excluded by σ. If all answers but one are excluded, an agent who is in state σ believes that one of the complete answers to the question is true: she knows a complete answer to the question. If none of the answers in Π are excluded, i.e. if σ overlaps with all blocks in the partition Π, this means that someone in state σ believes that all answers to the question are possible: we can say that σ is completely ignorant about Π.

The following definition specifies one way of specifying the information contained in a state about a certain partition.

Definition 3.36 Let Π be a partition and let σ be an information state. We define:

- The information contained in σ about Π, $\sigma \lhd \Pi$, is the set $\bigcup\{\pi \in \Pi \mid \pi \cap \sigma \neq \emptyset\}$.
- σ contains more information about Π than τ iff $\sigma \lhd \Pi \subseteq \tau \lhd \Pi$.
- σ contains complete information about Π iff $\sigma \lhd \Pi \in \Pi$. □

So, given an information state σ, the information about Π in σ is characterized by an information state τ that consists of the union of all complete answers in Π that are compatible with σ. We say that σ contains more information about Π if σ excludes more answers in Π than τ does. Finally, σ contains complete information about Π iff the information in σ about Π is a complete answer in Π.

We can see the function $\lhd\Pi$ that takes an information state to what is known in that information state as a filter on the information. Although the information state σ may contain more information, we are only interested in the information that σ contains about Π.

Extending these notions to our different notions of group knowledge is not very difficult.

Definition 3.37

- The information of a about Π in w is the information in $w(a)$ about Π.

- The common knowledge between $\mathcal{B}$ about Π in w, $C^{\Pi}_{\mathcal{B}}(w)$, is the information in $C_{\mathcal{B}}(w)$ about Π.

- The distributed knowledge of $\mathcal{B}$ about Π in w, $D^{\Pi}_{\mathcal{B}}(w)$, is the intersection of what each of the agents in $\mathcal{B}$ knows about Π, i.e.

$$D^{\Pi}_{\mathcal{B}}(w) = \bigcap_{a\in\mathcal{B}}(w(a)\lhd\Pi)$$

- The combined knowledge of $\mathcal{B}$ about Π, $I^{\Pi}_{\mathcal{B}}(w)$, is the set $\sigma = \{v \mid \exists u \in D^{\Pi}_{\mathcal{B}}(w) : u[\mathcal{B}]v$ and $v(b) = \sigma$ for each $b \in \mathcal{B}\}$ □

Also these notions can be visualised straightforwardly. There are three partitions pictured in figure 3.6. In each of the partitions, the surface surrounded by a dotted line represents an information state: it is a subset of the whole logical space. The double line surrounds the area that represents the information contained in the information about the partition. On the left hand side, the double line surrounds all blocks in the partition that are compatible with the information state. This is the information contained in the information state about the partition.

In the partition in the middle, the double line surrounds the blocks in the partition that are compatible with both information states. Note that in this picture, the distributed knowledge about the partition coincides with the information about the partition that is contained in the intersection of the two states. The right-most picture shows a case that is of most interest here. The intersection of the two states is empty, which means that the information distributed over the two states is inconsistent. But if we consider the information about the partition that is distributed over the two states, we see that there is one block compatible with both states. So, even if their distributed knowledge is inconsistent in an absolute sense, relative to this partition it is consistent.

If we have a language in which we can express questions, then we can also represent these notions in the object language. I will define a very simple extension of the language so we can express simple questions.

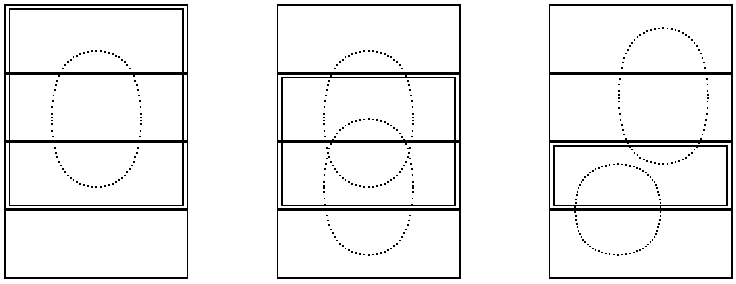


Figure 3.6.

Definition 3.38 (questions)
If ϕ is a sentence, then $¿\phi$ is a question, and if α and β are questions, then so is $\alpha \wedge \beta$ $\square$

Definition 3.39 (semantics of questions)
For each sentence ϕ, the partition $[\![¿\phi]\!]$ induced by the question $¿\phi$ consists of the two classes $\{w \mid w \models \phi\}$ and $\{w \mid w \models \neg\phi\}$.

If α and β are questions, then the partition induced by $\alpha \wedge \beta$ consists of all non-empty classes $\pi \cap \pi'$ such that $\pi \in [\![\alpha]\!]$ and $\pi' \in [\![\beta]\!]$. $\square$

For example, consider the question whether p. This question gives rise to a bipartition consisting of two blocks. The question $¿q$ does denotes a similar partition. We can combine the two questions to form a new question $¿p \wedge ¿q$, that denotes the partition consisting of four blocks, each block consisting of a class of possibilities where p and q have the same truth value. I have drawn these partitions together with another one in figure 3.7

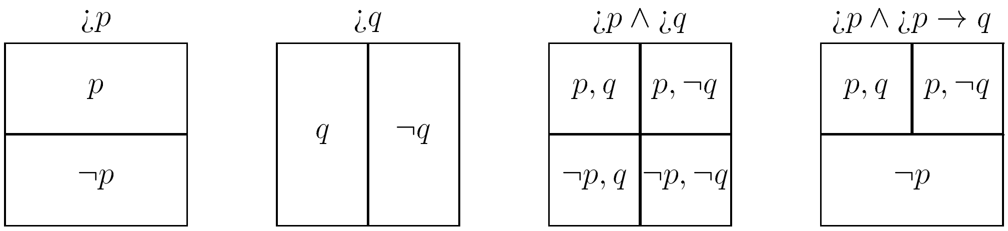


Figure 3.7. Some questions

Definition 3.40 For each question α and each subset $\mathcal{B}$ of agents, we add new operators $\square_a^\alpha$, $C_\mathcal{B}^\alpha$, $D_\mathcal{B}^\alpha$ and $I_\mathcal{B}^\alpha$ to the language.

Let α be a question, and define:

$$
\begin{array}{rll}
w \models \Box_b^\alpha \psi & \text{iff} & v \models \psi \text{ for each } v \in w(b) \lhd [\![\alpha]\!] \\
w \models C_{\mathcal{B}}^\alpha \psi & \text{iff} & v \models \psi \text{ for each } v \in C_{\mathcal{B}}^{[\![\alpha]\!]}(w) \\
w \models D_{\mathcal{B}}^\alpha \psi & \text{iff} & v \models \psi \text{ for each } v \in D_{\mathcal{B}}^{[\![\alpha]\!]}(w) \\
w \models I_{\mathcal{B}}^\alpha \psi & \text{iff} & v \models \psi \text{ for each } v \in I_{\mathcal{B}}^{[\![\alpha]\!]}(w)
\end{array}
\qquad \Box
$$

Sentences of the form $\Box_a^\alpha \psi$ should be read as: "a believes ψ in relation to the question α," sentences of the form $D_{\mathcal{B}}^\alpha \psi$ as "In relation to α, the sentence ψ is distributed knowledge between $\mathcal{B}$."

These relativised versions of epistemic operators can be useful in several ways. First of all, they provide a crude but simple way of talking about information about a topic without complicating the notion of an information state. Second, they give us a way of circumventing the inconsistency problem for distributed knowledge noted above.

For an example, consider an introspective possibility w for two agents such that $w(a)$ contains only worlds where p is true and $w(b)$ contains only worlds where p is false. Suppose also that a believes that q and that b believes that $q \to r$.

The definitions predict that the distributed knowledge of a and b about the question whether p is inconsistent, and that the distributed knowledge of the two agents about the questions whether q or r is consistent, and that about these questions, they mutually know that r is the case.

Let's first consider the information about p. In the partition associated with the question whether p is the case, all possibilities where p is true are grouped together, and all possibilities where p is false are grouped together. This means that the information state of a belongs to a different block in the partition than that of b. If we let $[\![¿p]\!]$ denote the partition associated with the question whether p, then $w(a) \lhd [\![¿p]\!]$ is the class of possibilities where p is true, and $w(b) \lhd [\![¿p]\!]$ consists of those possibilities where p is false. Clearly, the intersection of these two classes is empty, and therefore, their distributed knowledge about $[\![¿p]\!]$ is inconsistent.

In the partition associated with the intersection of the two questions whether q and whether r, all possibilities where q and r have the same truth-value are grouped together. The only block in this partition that overlaps with the information states of a as well as that of b is the block where both q and r are true. Relative to this partition, a and b distributively believe that r is true. Note that relative to this partition they have no distributed knowledge at all about the truth value of p.

If we consider the question whether r only, then the distributed knowledge of a and b is trivial, in the sense that only tautologies are distributed information about the question $¿r$. The reason for this is that neither of the two agents has any information about r, because neither one knows whether r is true or not. The agents need to be interested in q before they can use the fact that one of them knows q to conclude that r.

The 'absolute' notion of distributed knowledge is a limit case of the relativised one, where Π is the set of all singleton sets of possibilities. In that case, the notion of distributed knowledge in Π and that of the intersective definition 3.32 coincide.

Proposition 3.41 Let $\Pi = \{\{w\} \mid w \text{ is a possibility }\}$. Then the distributed knowledge of $\mathcal{B}$ in w in relation to Π is exactly the intersection $\bigcap_{a \in \mathcal{B}} w(a)$:

$$D^{\Pi}_{\mathcal{B}}(w) = D_{\mathcal{B}}(w)$$

The corresponding fact holds for Kripke models. $\Box$

So, we can see the concepts of knowledge and belief (and their mutual and distributed versions) as limit cases of knowledge about the most fine-grained question.

Note as an aside that in many cases (relative to partitions that are bisimulation-invariant), what information is distributed knowledge relative to a partition is independent of the fact whether we use Kripke models or possibilities as our semantics. The relativised versions of epistemic operators that we introduced in this section force us to be explicit about the differences between the elements of information states that we are interested in.

Logic

I will leave the question of a complete axiom system for the validities of the relativised semantics to another occasion, and only mention some interesting validities and invalidities of the logic.

First of all, the logic of $\Box^{\alpha}$ is not very different from that of $\Box$. Necessitation and distribution hold for all the operators $\Box^{\alpha}_{a}$: they are normal modal operators. If we look at the axioms specific for belief and knowledge, then also these are valid in their relativised versions.

If we consider the interaction between the different versions of $\Box^{\alpha}_{a}$, we need to use the notions of implication between questions of Groenendijk and Stokhof. In their semantics, a question implies another one just in case each complete answer to the first question implies a complete answer to the second. We write $\alpha \models \beta$ if α implies β.

If $\alpha \models \beta$, then $\vdash \Box^{\beta}_{a}\phi \rightarrow \Box^{\alpha}_{a}\phi$

The validity of the following axiom shows that the operators $\Box^{\alpha}_{a}$ are all weaker than the 'absolute' epistemic operator $\Box_{a}$.

$\vdash \Box^{\alpha}_{a}\phi \rightarrow \Box_{a}\phi$

However, when we consider positive introspection, the operator $\Box^{\alpha}_{a}$ behaves differently from $\Box_{a}$. For example, the law that $\Box^{\alpha}_{a}\phi \rightarrow \Box^{\alpha}_{a}\Box^{\alpha}_{a}\phi$ is not valid in introspective models. The point is that the premiss states that a believes that ϕ and that ϕ is relevant for α. The conclusion states that $\Box_a^\alpha \phi$ is relevant for α, which does not follow at all.

When we consider common knowledge, the situation is slightly more difficult. Again, the operators C^α are normal modal operators, but the logic is different from its absolute counterpart. In particular, laws such as $C^\alpha\phi \to C^\alpha C^\alpha \phi$ are not valid anymore.

The axiom and the rule for common knowledge need some small tinkering to work for the relativised version as well. The following versions are valid.

C1′ $C_{\mathcal{B}}^\alpha \phi \to (\Box_a^\alpha \phi \wedge \Box_a C_{\mathcal{B}}^\alpha \phi)$

RC′ If $\phi \to (\Box_a^\alpha \psi \wedge \Box_a \phi)$, then $\phi \to C_{\mathcal{B}}^\alpha \psi$

The rule $\mathsf{C1}'$ is slightly weaker than the straightforward version of $\mathsf{C1}$ in which all occurrences of epistemic operators are replaced by their relativised versions. The axiom states that each sentence ϕ that is about α and is common knowledge implies each of the agents knows ϕ about α, and knows that ϕ is common knowledge. They do *not* know that $C_{\mathcal{B}}^\alpha \phi$ about α, because the sentence $C_{\mathcal{B}}^\alpha \phi$ may not be relevant with respect to α at all. The rule RC' is slightly weaker than the version obtained by simply replacing the operators by their relativised versions, for similar reasons.

Also the relativised version of the operator for distributed knowledge is a normal modal operator (necessitation is true, and the operator distributed over implication). We can straightforwardly adapt the first two axioms:

D1′ $D_{\{a\}}^\alpha \phi \leftrightarrow \Box_a^\alpha \phi$

D2′ $D_{\mathcal{B}}^\alpha(\phi \to \psi) \to (D_{\mathcal{C}}^\alpha \phi \to D_{\mathcal{D}}^\alpha \psi)$ if $\mathcal{B} \subseteq \mathcal{C}$ and $\mathcal{C} \subseteq \mathcal{D}$

The relativised version of the knowledge axiom is also unproblematic: in reflexive models, the axiom is valid.

DT′ $D_{\mathcal{B}}^\alpha \phi \to \phi$

The relativised versions of positive and negative introspection of the distributed knowledge operators are not straightforwardly valid anymore. Consider for example the sentence $D_{\mathcal{B}}^\alpha \phi \to D_{\mathcal{B}}^\alpha D_{\mathcal{B}}^\alpha \phi$. There is no reason why this axiom should be valid: the fact that ϕ is distributed knowledge relative to α does not imply anything about the relevance of the sentence $D_{\mathcal{B}}^\alpha \phi$ to α. And if it is not relevant, then $D_{\mathcal{B}}^\alpha \phi$ is not distributed knowledge relative to α.

3.6 Conclusions and Open Questions

In this chapter I have studied classical multi-agent epistemic logic and some extensions of it. Looking back, a few themes can be discerned.

One important theme in this chapter is the role of ones ontological stance on possible worlds semantics when defining the semantics of epistemic operators. We saw already the difference between situation theory and possible worlds semantics when defining common knowledge. I have shown how different definitions give rise to the same truth-conditions in possibilities and Kripke models, but in situation theory, these different analyses give rise to different logics of common knowledge. With distributed knowledge and combined knowledge, the case was more subtle. We saw that a semantical definition of distributed knowledge depends directly on the way one interprets possible worlds semantics. In particular, we saw that the truth-conditions of sentences under the interpretation of D as intersection depends on the identity criteria between possible worlds. However, in the last section, where we considered information relative to a certain question, these problems seemed to dissolve.

Another important theme in this chapter is more technical, and concerns matters of completeness. We have seen that for classical logic, it is straightforward to adapt the standard Henkin style completeness proof to apply to possibilities. With respect to the semantics of distributed knowledge, the case was different, because the operator is not 'bisimulation invariant' (bisimilar models do not necessarily satisfy the same sentences). There are several questions that are still open. For example, the logic of distributed knowledge, defined as intersection in a language with a finite number of propositional variables is different from the logic DK, but I do not know what exactly the difference is. Another open question is what the logic of the full language $\mathcal{L}^{CDI}$ is, where the operators $\mathcal{I}_{\mathcal{B}}$ are allowed to occur with any non-empty subset of agents as a subscript. Also, there is the matter of completeness of the relativised versions of the epistemic operators. Finally, there is a small question whether DK45 and DS5 have the finite model property (it is already known that DK has be finite model property, and that the logics DK45 and DS5 are decidable).

When combined knowledge was discussed, I already touched upon its relation with information exchange. Information exchange between agents is the topic of the last chapter of this dissertation, and it is there that I will explore the notion a bit further. Before talking about information exchange in particular, we need to make sense of information change in general. This is the topic of the next chapter.

4

Dynamic Epistemic Semantics

In this chapter, I will finally turn to what I consider to be the most important contribution of this dissertation: the development of a 'dynamic epistemic semantics.'

Dynamic semantics is a branch of formal semantics that is concerned with *change*, and more in particular with change of information. The main idea in dynamic semantics is that the meaning of a syntactic unit—be it a sentence of natural language or a computer program—is best described as the change it brings about in the state of a human being or a computer. Motivation for, and applications of this idea can be found in areas such as semantics of programming languages (cf. Harel (1984)), default logic (Veltman (1996)), relevance logic (Belnap (1977)), pragmatics of natural language (Stalnaker (1972)), theory of anaphora (Groenendijk and Stokhof (1991), Kamp and Reyle (1993), Heim (1982)) and presupposition theory (Beaver (1995)). Van Benthem (1996) provides a mathematically oriented survey that contains many further references.

In this chapter some ideas from dynamic semantics are combined with the analysis of epistemic logic in terms of possible worlds semantics, as it was discussed in the previous chapter of this dissertation. I will develop a semantics and a deduction system for a multi-agent modal language extended with a repertoire of 'programs' (in the computer-related sense) that describe information change. The idea to extend multi-agent epistemic logic with operations that express information change is not new. In this chapter and the following, I will discuss the work of Landman (1986), Van Emde Boas et al. (1980), Van der Hoek et al. (1994b) and Groeneveld (1995), Jaspars (1994) and Fagin et al. (1995), who all did closely related work.

The chapter is organized as follows. In the next section, I will give a definition of Update Semantics (Veltman (1996)), which is at the basis of the dynamic logics of the rest of this chapter. In section 4.2 I try to make clear what a dynamic epistemic semantics should be able to do, and identify some potential problems. Section 4.3 contains the definition of Dynamic Epistemic Semantics proper. In the final section, I give a sound and complete axiomatization of the logic, and discuss some of the theorems and non-theorems of the logic.

The work in this chapter is based on the work in my master's thesis (Gerbrandy (1994)) and the articles Gerbrandy and Groeneveld (1997) and Gerbrandy (1997c).

4.1 Update Semantics

Update semantics (Veltman (1996)) models the change in information of an agent who learns sentences of propositional modal logic. It is also one of the logical systems that lies at the basis of the semantics that is the topic of this chapter. Here, I will introduce and discuss only the bare bones of the theory.

The language of Update Semantics is a standard propositional modal language.

Definition 4.1 (language of Update Semantics)
Let $\mathcal{P}$ be a set of propositional variables.

The set of *sentences* of Update Semantics, $\mathcal{L}^{\mathsf{US}}$, is the smallest set that contains $\mathcal{P}$, and is such that if ϕ and ψ are sentences of $\mathcal{L}^{\mathsf{US}}$, then so are $\neg\phi$, $\phi \wedge \psi$ and $\diamond\phi$.

If $\phi_0 \dots \phi_n$ are sentences of $\mathcal{L}^{\mathsf{US}}$, then $\phi_0; \phi_1; \dots; \phi_n$ is a *text* of $\mathcal{L}^{\mathsf{US}}$. □

The sentences $\mathcal{L}^{\mathsf{US}}$ are just the sentences of a classical modal language. New is the notion of a text, which is just a sequence of sentences.

In Update Semantics, the meaning of a sentence is identified with the effect it has on the information state of someone who accepts the information conveyed by that sentence. The notion of information state used in Update Semantics is a very simple one.

Definition 4.2 (information states)

- A possible world w assigns to each propositional variable a truth-value; it is a function $w : \mathcal{P} \mapsto \{0, 1\}$.
- An information state σ is a set of possible worlds. □

The intuition behind this notion of an information state is the same as that in classical epistemic logic: an information state of an agent is the set of 'possible ways the world might be that are compatible with the information of that agent:' for all the agent knows, each possible world in σ may picture reality correctly. For example, the information state consisting of the set of all possible worlds represents an information state of an agent that has no information about the world at all, and the empty set is an information state in which an agent has contradictory information.

Update Semantics is concerned with describing how the information state of an agent changes upon receiving information expressed by a sentence ϕ. Mathematically, we can model this effect by a function that, applied to an information state, returns a new information state that corresponds to the result of updating the first state with the information expressed by ϕ. When σ is an information state, and ϕ a sentence, we will write $\sigma[\phi]$ for the result of updating σ with the sentence ϕ. Intuitively, $\sigma[\phi]$ is the state that results when the agent, being in state σ, gets the information expressed by ϕ.

Definition 4.3 (interpretation)

$$\begin{array}{rcl}
\sigma[p] & = & \{w \in \sigma \mid w(p) = 1\} \\
\sigma[\neg\phi] & = & \sigma \setminus \sigma[\phi] \\
\sigma[\phi \wedge \psi] & = & \sigma[\phi] \cap \sigma[\psi] \\
\sigma[\Diamond\phi] & = & \begin{cases} \sigma & \text{if } \sigma[\phi] \neq \emptyset \\ \emptyset & \text{if } \sigma[\phi] = \emptyset \end{cases}
\end{array}$$

If $\phi_1; \ldots; \phi_n$ is a text, then:

$$\sigma[\phi_1; \ldots; \phi_n] \quad = \quad \sigma[\phi_1] \ldots [\phi_n]$$ □

An update of an information state σ with a sentence p results in a state containing all and only the worlds in σ in which p is true. The result of updating a state σ with a negated sentence $\neg\phi$ is a state containing all worlds in σ that do *not* survive in σ updated with ϕ. Conjunction is defined as intersection.

A sentence of the form $\Diamond\phi$ is interpreted as a consistency test on the information state of the hearer. One of two things may happen: either ϕ is consistent with the input state, in which case an update with $\Diamond\phi$ does not do anything at all, or ϕ is not consistent with the input state, in which case the update returns the empty (inconsistent) state. The idea is that this behavior corresponds with the behavior of constructions such as "It might be that ϕ."

Finally, the meaning of a text is simply the effect of processing all sentences of the text one by one, in the order that they are given.

Definition 4.4 (acceptance)

- A sentence ϕ is *accepted* in an information state σ iff $\sigma[\phi] = \sigma$. Notation: $\sigma \models \phi$. □

A sentence ϕ is accepted in an information state σ if an update with ϕ does not change the information state. Intuitively, this happens when the information expressed by ϕ is already contained in σ. If the information of an agent is represented by a state σ where ϕ is accepted, then that agent already believes that ϕ is true.

Update Semantics has some interesting extensions: in the original paper, Veltman develops a semantics for default reasoning based on Update Semantics, while in Groenendijk et al. (1996) Update Semantics is combined with Dynamic Predicate Logic, resulting in a semantics for the language of modal predicate logic in which anaphoric dependencies can be accounted for. Beaver (1995) extends the latter system with a (modal) operator that makes a sentence in the scope of it behave as a presupposition.

There is a precise sense in which updates always lead to an increase of information. An information state σ can be said to contain at least as much information as τ iff $\sigma \subseteq \tau$, i.e. just in case all possibilities that are excluded by τ are also excluded by σ. Under this 'information ordering,' an update of an information state σ with a sentence ϕ always results in a state that contains more (or the same amount of) information as the original state:

$$\sigma[\phi] \subseteq \sigma \text{ for all } \phi$$

This does not mean, however, that if a sentence is accepted in σ, it will still be accepted after we updated σ with new information. For example, suppose that σ is a state that contains possibilities in which p is true as well as possibilities in which p is false. Updating σ with the sentence $\Diamond p$ will have no effect on σ: in σ, $\Diamond p$ is accepted. Updating σ with $\neg p$ does have a considerable effect: all possibilities in which p is true will be discarded. In the resulting state $\sigma[\neg p]$, the sentence $\neg p$ is accepted; and therefore, $\Diamond p$ is rejected.

Another interesting feature of Update Semantics is that the effect of the update process is sensitive to the order in which sentences are processed. In general, it is not the case that $\sigma[\phi][\psi] = \sigma[\psi][\phi]$. To see this, consider the state σ again, in which both p and $\neg p$ are possible. Then $\sigma[\Diamond\neg p][p] \neq \emptyset$, but $\sigma[p][\Diamond\neg p] = \emptyset$.

This reflects the fact that changing the order of the sentences of a text can make a difference for the meaning of the text. Consider for example the following two 'texts':

Someone is knocking at the door... It might be John... It is Mary.

Someone is knocking at the door... It is Mary... It might be John.

The first sequence of sentences is acceptable, while the second is not: once one knows it is Mary who is knocking at the door, it cannot be John anymore. This is reflected in Update Semantics: $\Diamond p; \neg p$ is consistent (in the sense that a contradiction is not derivable from the text) while $\neg p; \Diamond p$ is not.

This concludes my description of Update Semantics. In section 5.5, I will use dynamic epistemic semantics to give an analysis of Update Semantics that shows that Update Semantics can be seen as the logic of information change of a single agent in an introspective state.

4.2 Dynamic Epistemic Semantics

I will now turn to the project of combining Update Semantics with the multi-agent epistemic logics of the previous chapter. I will define operations on possibilities that correspond to changes in the information states of the agents. The kind of information change we will model is that of agents getting new information, similar to the kind of information change modeled in Update Semantics. In addition, I will also allow for the possibility that agents learn about each other's change of information, such as agent a learning that agent b has learned that ϕ.

To get an idea of what this involves, consider a simple example. Let the possibilities w and v be given by the following equations:

$$\begin{array}{rcl} w(p) & = & 1 \\ w(a) & = & w(b) = \{w, v\} \\ v(p) & = & 0 \\ v(a) & = & v(b) = \{w, v\} \end{array}$$

As can easily be checked, both w and v are introspective and truthful. This means that there is a picture of w in the form of a Kripke model with an accessibility relation that is an equivalence relation. Figure 4.1 is an example of such a picture.

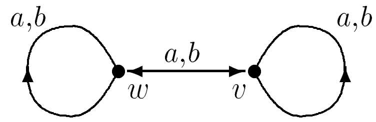


Figure 4.1. The possibilities w and v

Suppose that in the situation described by w, the agent a learns that p is the case. In Update Semantics, the effect of this is that a discards all possibilities in which p is false from her information state. Let us also assume that b is not aware that a is getting new information, i.e. that b's information state is not affected. The resulting possibility u would be a possibility that only differs from w in that in the new possibility u, a's information state $u(a)$ consists of all and only possibilities in $w(a)$ in which p is true. We will see later that this is not quite right, but let us stick to this simple picture for the moment. The resulting possibility u is given by the following equations:

$$u(p) \quad = \quad w(p)$$

$$\begin{array}{rcl} u(b) & = & w(b) \\ u(a) & = & w(a) \setminus \{v\} = \{w\} \end{array}$$

We can draw a picture of u as in figure 4.2. But u is not the right possibility to

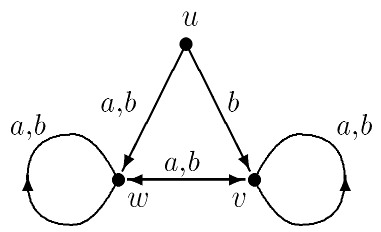


Figure 4.2. A picture of u

result after a learns that p in w. The point is that a's information is not factual in u: she does not consider the world u itself as a possible state of affairs. This cannot be right: first, a did not believe any falsehoods, then she learned that p, which was in fact true, so the result should be a state in which her beliefs are true as well.

To see what is going wrong, observe that u is a state in which a believes that p is the case, which is what we wanted, but it is also a state in which a believes that she does not know that p (in 'all' of a's alternatives in u, namely in w, a does not know that p). For a correct notion of belief update, we also need to take into account that after a learns that p, she not just believes that p, but also believes that she believes that p, etcetera. In other words, we have to find a notion of update for which it holds that if a learns that p, she learns that she has learned that p as well. Following terminology from Groeneveld (1995), I will call such an update a *conscious* update.

Consider the following definition of w_1:

$$\begin{array}{rcl} w_1(p) & = & w(p) \\ w_1(b) & = & w(b) \\ w_1(a) & = & \{w_1\} \end{array}$$

I believe that w_1 is the possibility that is the best description of the situation that is the result of a learning p in w. The possibility w_1 only differs from w in that the information state of a has changed. If a learns that p is the case, this means that she discards v from her information state, leaving only w as an option. But a also learns that she learns that p, i.e. that w has changed. The possibility that is the result when a learns that p in w is exactly the possibility w_1 that we are defining.

Figure 4.3 is a picture of w_1. Note that in w_1, a's information is introspective

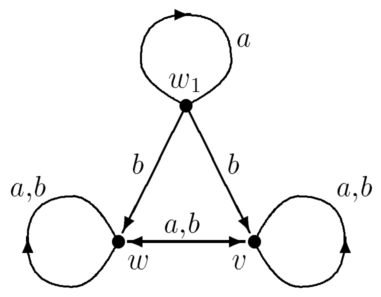


Figure 4.3. w_1 is the result of a learning that p in w

and factual, but b's information is, although introspective, not factual. This is what we wanted. Note moreover that a believes that p is the case, believes that she believes that p is the case, etcetera. Finally, observe that this is in a sense the *only* thing that a has learned: the information that a had about the information of b is unaffected.

Let us consider another, more complicated example. Suppose that in w_1, b learns that a has learned that p is the case. The resulting state, call it w_2, should be an introspective state in which b believes that a believes that p. In addition, the information in w_2 of b should be correct: since his information state was factual in w, and did not change when a learned that p, it should be correct in w_2, where he learned about the changes in a's information state that in fact took place.

On the other hand, in w_2, the information of a will *not* be correct, because in w_1, she believes that b believes that a does not know whether p, and her information will not be affected because b learned something new. So, in w_2, a will believe, falsely, that b believes that a does not know whether p.

Let us now look at the formal implementation of these ideas. In w_1, b has two belief alternatives, w and v. This means he does not know which of these two represents the real world. When he learns that a has learned that p, for all b knows, this may have happened in w as well as in v. So, a good way of modeling the effect of this update is to update both w and v with the information that a learns that p. We already know what the effect of this is on w: the result is w_1. The effect of a learning that p on v is similar; the resulting state v_1 will be just like w_1, except that p is false in v_1.

But, just as in the previous example, this is not enough; we also have to make sure that the update is conscious, that b also learns that he learns that p.

The resulting state w_2 will be one that differs from w_1 only in the information state that is assigned to b. In w_2, b's new state $w_2(b)$ will consist of two possibilities: the update of w with the information that a has learned that p, and that b knows that this change has occurred, and the update of v with the same information. Let us denote the respective results by w' and v'. These possibilities should satisfy the following equations:

$$\begin{array}{rcl} w'(p) & = & w(p) = 1 \\ w'(a) & = & w_1(a) = \{w_1\} \\ w'(b) & = & \{w', v'\} \end{array}$$

$$\begin{array}{rcl} v'(p) & = & v(p) = 0 \\ v'(a) & = & w_1(a) = \{w_1\} \\ v'(b) & = & \{w', v'\} \end{array}$$

In w_2, which is the result after b learns that a has learned that p in w_1, the new state of b will consist of the two possibilities w' and v':

$$\begin{array}{rcl} w_2(p) & = & w_1(p) = 1 \\ w_2(a) & = & w_1(a) = \{w_1\} \\ w_2(b) & = & \{w', v'\} \end{array}$$

These possibilities are represented in figure 4.4.

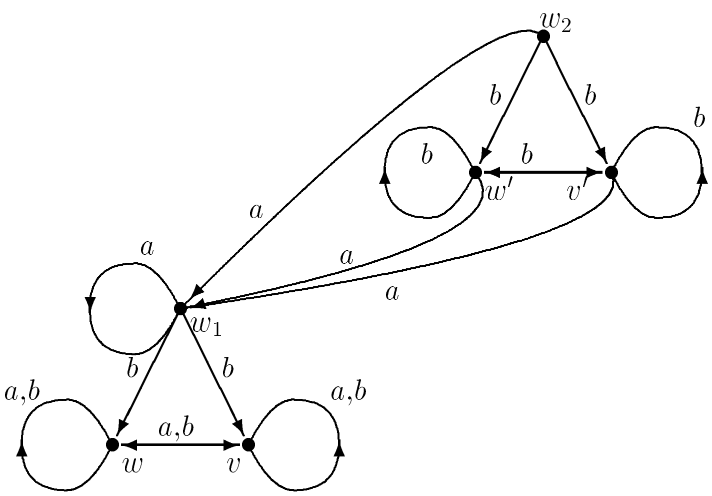


Figure 4.4. w_2 is the possibility after b learns that a has learned that p in w_1

This is not an extensional picture: the nodes w' and w_2 represent the same possibilities, since they assign the same values to propositional variables and assign the same states to both a and b. To see that the information of b is factual, i.e. that the 'actual world' is one of the epistemic alternatives of b, we can draw another picture of w_2 in which each possibility corresponds to a unique

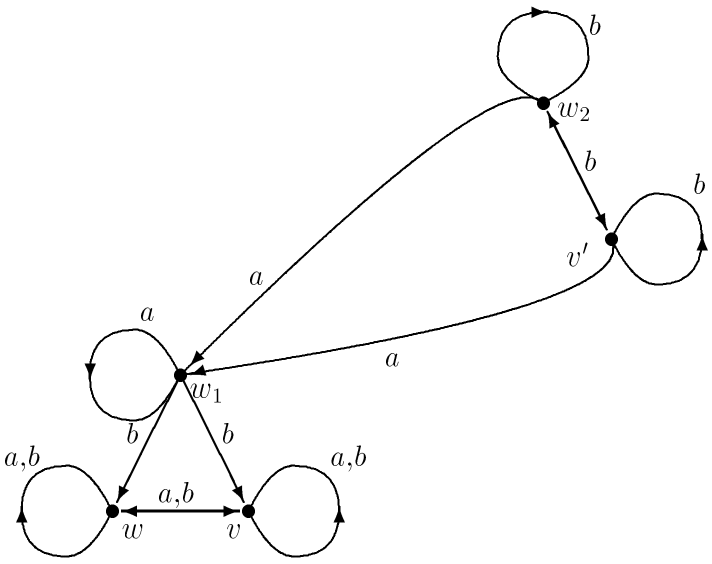


Figure 4.5. An extensional picture of w_2

node in the Kripke model. I have done this in figure 4.5. Note that a's beliefs are *not* factual in w_2: she believes something false, namely that b believes that she does not know that p.

Representing information change in extensional pictures, as I have done in the examples above, is rather awkward: the 'mechanics' of an update are not represented in a particularly clear way. A more perspicuous way of visualizing a conscious update is by drawing possibilities as graphs that look like a tree. In that case, we can see the examples of conscious updates above correspond to 'pruning' the tree: taking away certain branches. Let us consider our first example again of the possibility w that was pictured in figure 4.1. If we draw this model as a tree, we get the picture of figure 4.6.

To avoid overcrowding the picture with labels for agents, propositions and possibilities, I have represented the accessibility relations for a by continuous arrows, and those of b with broken lines. A node that is represented by a filled circle is a world in which p is true, a node represented by an open circle is one in which p false. In figure 4.6, the possibility w from figure 4.1 is represented as the top node of the tree. Note that the two pictures are bisimilar. A conscious update of a's information with p corresponds with removing all branches that can be reached by a-arrows from the top node in which p is false. This can be seen when we draw w_1 as a tree: the resulting picture is represented in figure 4.7. This picture differs from that of figure 4.6 in that all nodes are removed where p is false that can be reached from w by a sequence of a-transitions.

Our second example was the situation where b learns that a has learned that

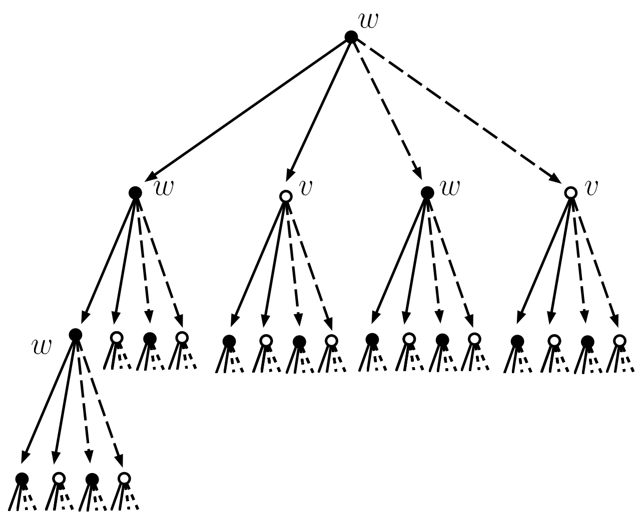


Figure 4.6. w as a tree.

p. As an update on trees, this corresponds to removing all branches in which p is false that can be reached by a sequence of b-arrows followed by a sequence of a-arrows. The result of doing this in the tree model that represents w_1 leads to a picture such as that of figure 4.8.

In the next section, I will provide precise definitions of operations of information change of possibilities. The idea that updates can be seen as operations on trees is worked out in detail in section 5.2.2.

4.3 Programs

We are now ready to start with the formal definitions of dynamic epistemic logic. First we need a language to describe the changes that may occur in the information states of agents. As said before, I will not only consider changes that are the result of agents learning certain facts about the world, but also the changes that are the result of agents learning about the information change of other agents.

To express change, I will introduce 'programs' in the object language. These programs will be interpreted as relations between possibilities: if π is a program, then its interpretation $[\![\pi]\!]$ will be that relation between possibilities that holds between two possibilities w and v just in case v is a possible output of executing π in w.

The most simple programs are of the form $?\phi$, where ϕ is a sentence. The program $?\phi$ is interpreted as a test: it can do nothing but succeed or fail. The test $?\phi$ succeeds in a possibility w exactly when ϕ is true, and will return w as its output. Otherwise it fails, and it will have no output at all.

All other programs are built from tests using certain operators. The most

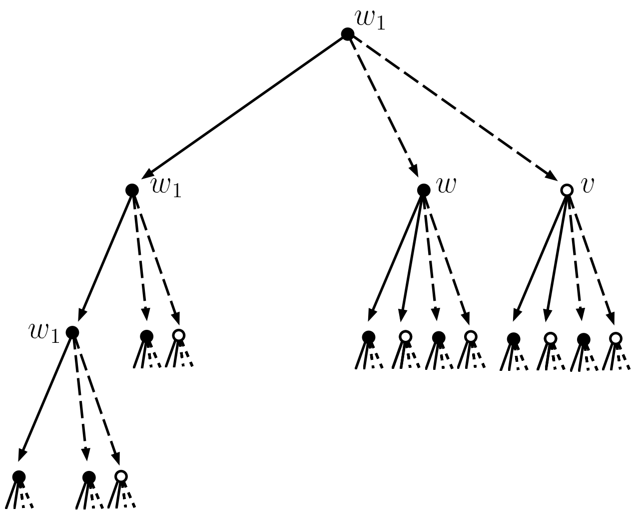


Figure 4.7. A picture of the possibility w_1 as a tree

important of these is written as U_a: if π is a program, then $U_a\pi$ expresses that a (consciously) learns that π has been (successfully) executed; I will often say that a updates with π. The kind of action discussed in the example above where a learns that p can be expressed in the language by the program $U_a?p$. If a learns that the test $?p$ has been successfully executed, then a will perform the test $?p$ at each of her epistemic alternatives. For those alternatives in which p is true, the test will change nothing at all, in those alternatives in which p is false, the test will fail, and return no output at all. The result is that all possibilities where the test fails are eliminated from the state of a. A further effect of an update with $U_a?p$ is that all remaining possibilities are updated with $U_b?p$ as well. The net result is that a consciously learns that p is true.

The second example, of b learning that a learned that p, can be expressed by the program $U_b(U_a?p)$.

Finally, the language contains two operators that combine programs to form a new program: sequencing and choice. A program of the form $\pi;\pi'$ is interpreted as "first execute π, then π'." The program $\pi \cup \pi'$ corresponds to executing either π or π'.

To connect the programming language to the 'static' language of multi-modal logic, we add a modal operator $[\pi]$ for each program π to the language. Intuitively, a sentence $[\pi]\psi$ is true in a possibility iff after executing the program π in that possibility, ψ is always true. The set of programs is defined simultaneously with the set of sentences. The syntax of the language is a variation of that of propositional dynamic logic (as in Pratt (1976) and Goldblatt (1987)).

Definition 4.5 (language of DEL)
Given a set of agents $\mathcal{A}$ and a set of propositional variables $\mathcal{P}$, we define the

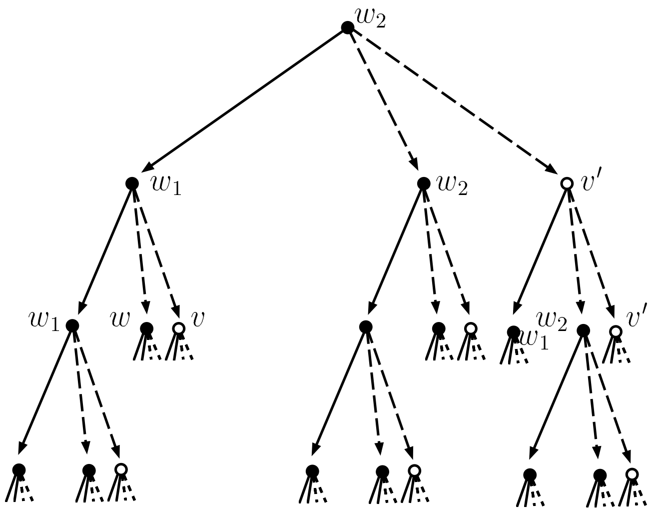


Figure 4.8. The possibility w_2

sentences and programs of DEL simultaneously as follows.

The set of *sentences* of DEL is the smallest set that contains $\mathcal{P}$, and such that if ϕ and ψ are sentences and π is a program, then $\neg\phi$, $\phi \wedge \psi$, $\Box_a\phi$ and $[\pi]\phi$ are sentences of DEL.

The set of *programs* is the smallest set that contains $?\phi$ for each sentence ϕ, and for which it holds that if π and π' are programs and a an agent, then $U_a\pi$, $\pi;\pi'$ and $\pi \cup \pi'$ are programs as well. □

The programming language is constructed in such a way that each program can be executed by each of the agents. This has the effect that any change in the model that we can express as a program in the object language can be 'learned' by each of the agents. In particular, because there is a test $?\phi$ in the programming language for each sentence ϕ, each sentence can be 'learned' by each of the agents,

Sentences will be interpreted as in static epistemic logic, i.e. we will define when a sentence is true in a possibility. Programs are interpreted as relations over possibilities: a pair of possibilities (w, v) will be in the denotation of a program π just in case the execution of the program π in possibility w may have v as its output. Instead of writing $(w, v) \in [\![\pi]\!]$, I will write $w[\![\pi]\!]v$ to express that w and v stand in the relation $[\![\pi]\!]$. Also, I use the abbreviation $w[a]v$ for the statement that w differs at most from v in the state it assigns to a:

Definition 4.6 (interpretation of programs)

$$\begin{array}{rll} w[\![?\phi]\!]v & \text{iff} & w \models \phi \text{ and } w = v \\ w[\![U_a\pi]\!]v & \text{iff} & w[a]v \text{ and } v(a) = \{v' \mid \exists w' \in w(a) \exists u : w'[\![\pi]\!]u[\![U_a\pi]\!]v'\} \\ w[\![\pi;\pi']\!]v & \text{iff} & \text{there is a } u \text{ such that } w[\![\pi]\!]u[\![\pi']\!]v \\ w[\![\pi \cup \pi']\!]v & \text{iff} & \text{either } w[\![\pi]\!]v \text{ or } w[\![\pi']\!]v \end{array}$$

Furthermore, the definition of truth (definition 3.5) is extended with the following clause:

$$w \models [\pi]\phi \quad \text{iff} \quad \text{for all } v \text{ if } w[\![\pi]\!]v \text{ then } v \models \phi$$

The other static operators get their standard interpretation. □

Programs of the form $U_a\pi$ are to be read as "a learns that π has been successfully executed," or, alternatively, as "a updates her information state with π." This is modeled as follows. Executing a program of the form $U_a\pi$ in a possibility w results in a new possibility v in which only a's information state has changed. The information state of a in v contains all and only those possibilities that are the possible result of a successful execution of π in one of the possibilities in a's information state in w. Since we want updates to be conscious, we also change the resulting possibilities to the effect that a has updated with π. So, the state of a after an update with $U_a\pi$ contains exactly those possibilities v' that are the result of updating a possibility w' from the old state first with π, and then with $U_a\pi$.

The definition of $[\![U_a\pi]\!]$ is circular as it stands: it is defined in terms of itself. Nevertheless, it is not very hard to prove that for each program π there is a unique relation $[\![U_a\pi]\!]$ that conforms to the definition. This follows from proposition 1.25. The function $[\![U_a\pi]\!]$ is defined along the format of the function $\phi_{\mathcal{B}}$ on page 17. A direct proof of the correctness of a similar definition can be found in Gerbrandy and Groeneveld (1997).

A program of the form $U_a\pi$ is deterministic.[1] In fact, $[\![U_a\pi]\!]$ is a total function: it always has an output. It holds that if a's information in w is introspective, it is introspective after the update of w with $U_a?\phi$, which gives evidence that $U_a\pi$ indeed has the effect that a conscious update should have.

It is not hard to check that these definitions define exactly the updates that we discussed in the previous section. It holds that $w[\![U_a?p]\!]w_1$ and that $w_1[\![U_bU_a?p]\!]w_2$, where w, w_1 and w_2 are as in the previous section.

We can also express more complicated changes. For example the program $U_b(?\top \cup U_a?p)$ expresses that b learns that either nothing happened, or that a learned that p. This is a deterministic program, expressing a certain change in the information of b. Its effect is very different from that of the program $U_b?\top \cup U_bU_a?p$, which is a non-deterministic program that expresses that either b learns that $\top$ is the case, or that b learns that a learns that p.

For another example, consider the program $(?p; U_a?p) \cup (?\neg p; U_a?\neg p)$. The effect of this program is that a updates with $?p$ if the test $?p$ succeeds, and that a updates with $?\neg p$ when $?\neg p$ succeeds. In other words, a learns that p if p is in fact true, and she learns that p is false if p is in fact false. This program can

[1] A program is deterministic iff it always has a unique output, i.e. when its meaning is a function.

be seen as expressing the effect of a learning whether p is the case. This may be useful to describe a situation, for example, of an agent a examining the value of a bit (p expressing that the value is 0, $\neg p$ expressing it is 1) or of a philosopher looking out of the window to check whether it rains, when we (the modellers) don't know what the value in fact is, or whether it rains or not.

Group Updates

The concept of common knowledge has been extensively discussed in chapter 3. The usual definition is that a sentence ϕ is common knowledge in a group $\mathcal{B}$ just in case each agent in the group knows that ϕ is the case, each agent knows that each of the other agents knows that ϕ, etcetera. For certain applications of DEL, it would be useful to be able to describe 'mutual information change.' For example, in some models of dialogue, the effect of an utterance of a sentence p is that p becomes common knowledge (cf. chapter 6.5). We cannot express such an update in our programming language as we have defined it so far, because the update operators U_a affect the state of one agent at a time. This observation can be backed up with the fact that as long as the information of a and of b remains true after an update with a program π, the information that is common knowledge does not change as a result of the update.

Proposition 4.7 For any reflexive w and each program π, it holds that if $w \not\models C_{\{a,b\}}\phi$ and $w[\![\pi]\!]v$ then $v \not\models C_{\{a,b\}}\phi$. □

This means that we cannot express in our language that certain information becomes common knowledge. This is an interesting fact in itself: it shows that group knowledge cannot change as a result of agents separately changing their information states, not even when they learn the same things. Changes in common knowledge are the result of something different: of agents *mutually* learning a certain fact. To model change in common knowledge, I will introduce the notion of a 'group update': an update with a program π in a group of agents has the effect of changing the state of each agent in the group in the way described by π, but in such a way that all agents involved are conscious of the fact that the whole group has performed the update. Roughly, the effect of a mutual update in a group $\mathcal{B}$ with π is that each agent in $\mathcal{B}$ updates with π, each agent updates with the information that each agent updated her state with π, etcetera.

To express this in the object language, I add program operators of the form $U_\mathcal{B}$ for each subset $\mathcal{B}$ of $\mathcal{A}$ to the language. They are interpreted as follows:

Definition 4.8 (group updates)
For each π and $\mathcal{B} \subseteq \mathcal{A}$:

$$\begin{array}{lll} w[\![U_\mathcal{B}\pi]\!]v & \text{iff} & w[\mathcal{B}]v \text{ and } \forall a \in \mathcal{B} : \\ & & v(a) = \{v' \mid \exists w' \in w(a) : w'[\![\pi]\!][\![U_\mathcal{B}\pi]\!]v'\} \end{array}$$

□

Updating a possibility w with a program $U_{\mathcal{B}}\pi$ results in a possibility v that differs only from w in that for each $a \in \mathcal{B}$, all possibilities in $w(a)$ are first updated with π, and then with $U_{\mathcal{B}}\pi$.

If the group is a singleton set, the effect of a group update is exactly the same as that of a conscious update by a single agent: updating with $U_{\{a\}}\pi$ has the same effect as updating with $U_a\pi$.

4.4 Axiomatization

Ideally, an axiom system for the kind of semantics I have defined in the previous section consists of two more or less independent modules. The language is divided into two interrelated parts, a set of programs and a set of sentences, and the semantics consists of separate definitions of the meaning of programs and of sentences. A logic for the semantics should consist of two parts as well: one part that axiomatizes the set of valid sentences, just as in static epistemic logic, and an 'algebraic' part that tells us when two programs have the same meaning. I will only give the former in this dissertation, for the simple reason that I have not been able to find an axiomatization of the identities between programs.

The following set of axioms and rules provides a sound and complete characterization of the set of sentences true in all possibilities. (For sake of presentation, I have left out the conscious single agent updates, since they are a special case of the group updates with a group consisting of a single agent.)

Axioms

1 $\vdash \phi$ if ϕ is valid in classical propositional logic.

2 $\vdash \Box_a(\phi \to \psi) \to (\Box_a\phi \to \Box_a\psi)$

3 $\vdash [\pi](\phi \to \psi) \to ([\pi]\phi \to [\pi]\psi)$.

4 $\vdash [?\phi]\psi \leftrightarrow (\phi \to \psi)$

5 $\vdash \neg[U_{\mathcal{B}}\pi]\psi \to [U_{\mathcal{B}}\pi]\neg\psi$ (functionality)[2]

6 $\vdash [U_{\mathcal{B}}\pi]p \leftrightarrow p$

7 $\vdash [U_{\mathcal{B}}\pi]\Box_a\phi \leftrightarrow \Box_a[\pi][U_{\mathcal{B}}\pi]\phi$ if $a \in \mathcal{B}$ (Ramsey axiom)

8 $\vdash [U_{\mathcal{B}}\pi]\Box_a\phi \leftrightarrow \Box_a\phi$ if $a \notin \mathcal{B}$ (privacy)

9 $\vdash [\pi;\pi']\phi \leftrightarrow [\pi][\pi']\phi$

10 $\vdash [\pi \cup \pi']\psi \leftrightarrow ([\pi]\psi \wedge [\pi']\psi)$

Rules

[2]The converse direction is derivable: $[U_{\mathcal{B}}\pi]\neg\psi \leftrightarrow [U_{\mathcal{B}}\pi](\psi \to \bot) \leftrightarrow$ (with axiom 3) $([U_{\mathcal{B}}\pi]\psi \to [U_{\mathcal{B}}\pi]\bot) \leftrightarrow$ (with axiom 6) $(([U_{\mathcal{B}}\pi]\psi) \to \bot)$

MP $\phi, \phi \to \psi \vdash \phi$

Nec□ If $\vdash \phi$ then $\vdash \Box_a \phi$

We write $\Gamma \vdash \phi$ or $\Gamma \vdash_{\mathsf{DELK}} \phi$ just in case there is derivation of ϕ from the premises in Γ using the rules and axioms above. The logics DELK45 and DELS5 are obtained from DEL by adding the extra axioms of K45 and S5 respectively, and we use the notation $\vdash_{\mathsf{DELK45}}$ and $\vdash_{\mathsf{DELS5}}$ accordingly. □

In addition to the rules and axioms of classical modal logic, the deduction system consists of axioms describing the behavior of the program operators. Axiom 3 tells us that the set of sentences true after an update is closed under logical consequence. Axiom 4 says that performing a test $?\phi$ boils down to checking whether ϕ is true: if ψ is true after each successful execution of the test $?\phi$, then it is true that ϕ implies ψ. Vice versa, if ϕ implies that ψ, and the test $?\phi$ succeeds, then ψ must be true. The functionality axiom 5 reflects the fact that U_a-updates are total functions: an update with $U_a\pi$ always gives a unique result. This means that if it is not the case that a certain sentence is true after an update with a program of the form $U_{\mathcal{B}}\pi$, then it must be the case that the negation of that sentence is true in the updated possibility. Axiom 6 expresses that the update of an information state has no effect on the 'real' world; the same propositional atoms will be true or false before and after an update. The Ramsey axiom, axiom 7, expresses that after a group update with π, an agent in the group knows that ψ just in case he already knew that after executing π, an update with $U_{\mathcal{B}}\pi$ could only result in a possibility in which ψ were true.

The privacy axiom 8 expresses that a group update has no effect on the information of agents outside that group. The axioms 9 and 10 govern the behavior of sequencing and disjunction respectively.

I will discuss the validities of this axiom system in more detail later, and first prove that the axiom system given above is sound and complete. The following proposition shows that the axioms are sound: if some sentence is derivable, it must be true in all models.

Proposition 4.9 (soundness)
If $\Gamma \vdash_{\mathsf{DELK}} \phi$ then $\Gamma \models \phi$
If $\Gamma \vdash_{\mathsf{DELK45}} \phi$ then $\Gamma \models_{\mathsf{belief}} \phi$
If $\Gamma \vdash_{\mathsf{DELS5}} \phi$ then $\Gamma \models_{\mathsf{knowledge}} \phi$

proof: By a standard induction. By way of illustration, I will demonstrate that axiom 5 is sound. In the proof, I make use of the fact that $[\![U_{\mathcal{B}}\pi]\!]$ is a total function for each π, and write $w[\![U_{\mathcal{B}}\pi]\!]$ for the unique v such that $w[\![U_{\mathcal{B}}\pi]\!]v$. The following equivalences hold, if $a \in \mathcal{B}$:

$$\begin{array}{lll} w \models [U_{\mathcal{B}}\pi]\Box_a\phi & \text{iff} & w[\![U_{\mathcal{B}}\pi]\!] \models \Box_a\phi \\ & \text{iff} & \forall v \in w[\![U_{\mathcal{B}}\pi]\!](a) : v \models \phi \end{array}$$

iff $\forall v :$ if $\exists w' \in w(a) : w' [\![\pi]\!] [\![U_{\mathcal{B}}\pi]\!] v$ then $v \models \phi$
iff $\forall w' \in w(a)\, \forall v :$ if $w' [\![\pi]\!] [\![U_{\mathcal{B}}\pi]\!] v$ then $v \models \phi$
iff $\forall w' \in w(a) : w' \models [\pi][U_{\mathcal{B}}\pi]\phi$
iff $w \models \Box_a[\pi][U_{\mathcal{B}}\pi]\phi$

Checking the other axioms for soundness is comparatively easy. □

The axiom system is complete as well: all valid sentences can be derived in the logic.

Proposition 4.10 (completeness)
If $\Gamma \models \phi$, then $\Gamma \vdash \phi$
If $\Gamma \models_{\mathsf{belief}} \phi$, then $\Gamma \vdash_{\mathsf{DELK45}} \phi$
If $\Gamma \models_{\mathsf{knowledge}} \phi$, then $\Gamma \vdash_{\mathsf{DELS5}} \phi$

proof: In Gerbrandy (1997c) and Gerbrandy and Groeneveld (1997), an axiom system similar to DELK was proven to be complete with the help of the Henkin method. In the proof we used a necessitation rule for program modalities. For reasons that I will give later, I have not added such a rule to this version of DEL, which has a consequence that we cannot use the earlier completeness proof.

In this section I will give a proof of completeness that is less direct. First, I show in proposition 4.13 that each sentence of DEL is equivalent to one without any program modalities. Completeness of the logic is then a direct consequence of the completeness of the classical logics K, S5 and K45. □

Consider the following definition:

Definition 4.11 (elimination of programs)
The function $(\cdot)^*$ is given by:

1. $(p)^* = p$
2. $(\phi \wedge \psi)^* = (\phi)^* \wedge (\psi)^*$
3. $(\neg\phi)^* = \neg(\phi)^*$
4. $(\Box_a\phi)^* = \Box_a(\phi)^*$
5. $([?\phi]\psi)^* = (\phi)^* \rightarrow (\psi)^*$
6. $([U_{\mathcal{B}}\pi]p)^* = p$
7. $([U_{\mathcal{B}}\pi](\phi \wedge \psi))^* = ([U_{\mathcal{B}}\pi]\phi)^* \wedge ([U_{\mathcal{B}}\pi]\psi)^*$
8. $([U_{\mathcal{B}}\pi]\neg\phi)^* = \neg([U_{\mathcal{B}}\pi]\phi)^*$
9. $([U_{\mathcal{B}}\pi]\Box_a\phi)^* = \Box_a(\phi)^*$ if $a \notin \mathcal{B}$
10. $([U_{\mathcal{B}}\pi]\Box_a\phi)^* = \Box_a([\pi]([U_{\mathcal{B}}\pi]\phi)^*)^*$ if $a \in \mathcal{B}$
11. $([U_{\mathcal{B}}\pi][\pi']\psi)^* = ([U_{\mathcal{B}}\pi]([\pi']\psi)^*)^*$

12. $([\pi \cup \pi']\phi)^* = ([\pi]\phi)^* \wedge ([\pi']\phi)^*$
13. $([\pi;\pi']\phi)^* = ([\pi]([\pi']\phi)^*)^*$ □

This is a rather complicated definition. Given clause 10, it may not even be immediately clear that this definition is a correct recursive one. To see that it is, consider the following definition of the complexity of sentences of DEL.

Definition 4.12 (complexity)

- 'Standard' complexity sc:
  $\mathsf{sc}(p) = 1$
  $\mathsf{sc}(\phi \wedge \psi) = \max(\mathsf{sc}(\phi), \mathsf{sc}(\psi)) + 1$
  $\mathsf{sc}(\neg\phi) = \mathsf{sc}(\Box_a\phi) = \mathsf{sc}(\phi) + 1$
  $\mathsf{sc}([\pi]\phi) = \mathsf{pc}(\pi) + \mathsf{sc}(\phi)$

- Program complexity pc:

$\mathsf{pc}(p) = 0$	$\mathsf{pc}(?\phi) = \mathsf{pc}(\phi) + 1$
$\mathsf{pc}(\phi \wedge \psi) = \max(\mathsf{pc}(\phi), \mathsf{pc}(\psi))$	$\mathsf{pc}(U_{\mathcal{B}}\pi) = \mathsf{pc}(\pi) + 1$
$\mathsf{pc}(\neg\phi) = \mathsf{pc}(\Box_a\phi) = \mathsf{pc}(\phi)$	$\mathsf{pc}(\pi \cup \pi') =$
$\mathsf{pc}([\pi]\phi) = (\mathsf{pc}(\pi) + \mathsf{pc}(\phi))$	$\mathsf{pc}(\pi;\pi') = \max(\mathsf{pc}(\pi), \mathsf{pc}(\pi')) + 1$

- A sentence ϕ is less complex than a sentence ψ iff either $\mathsf{pc}(\phi) < \mathsf{pc}(\psi)$ or $\mathsf{pc}(\phi) = \mathsf{pc}(\psi)$ and $\mathsf{sc}(\phi) < \mathsf{sc}(\psi)$. □

The function sc measures the complexity of a sentence in more or less the standard way by counting the nesting of logical operators in a sentence, while the program complexity of a sentence is a measure of the amount of nested programs occurring in that sentence.

The general notion of complexity we will use is defined on the basis of these two notions. In the complexity ordering, program complexity takes precedence over 'normal' complexity; i.e. ϕ is more complex than ψ iff it either contains 'more programs', or, if they both have the same program complexity, we use the standard notion of complexity to measure the difference.

We can now show that the translation function of definition 4.11 is a recursive definition, in the sense that the definition of the value of $(\phi)^*$ makes use of the value of $(\psi)^*$ only if ψ is of lower complexity than ϕ. I will also use the same complexity ordering to give an inductive proof of the fact that $(\phi)^*$ does not contain any program operators, and that ϕ and $(\phi)^*$ are equivalent in DEL.

Proposition 4.13
(1) Definition 4.11 is a correct recursive definition
(2) ϕ^* is a sentence of classical modal logic.
(3) ϕ and ϕ^* are provably equivalent in DEL.

proof: I will prove (1), (2) and (3) at the same time, by induction on the complexity of ϕ. More precisely, the proof follows the clauses in the definition of $(\cdot)^*$, but the induction hypothesis is taken to hold for all sentences of lower complexity. In all three proofs, the difficult cases are that of clause 10, 11 and 13, in which there are two occurrences of * on the right side.

To prove (1), we need to show that in each clause in the definition of $(\cdot)^*$, the sentences on the right hand side to which the function $(\cdot)^*$ is applied are of lower complexity than the sentence on the left hand side. The result of (2) follows almost directly from this, while (3) is easy to prove: almost each clause in the definition of $(\cdot)^*$ corresponds to an axiom.

Just to illustrate, consider clause 10 of the translation function, which says that if $a \in \mathcal{B}$, then $([U_{\mathcal{B}}\pi]\Box_a\phi)^* = \Box_a([\pi]([U_{\mathcal{B}}\pi]\phi)^*)^*$.

We need to show three things. First, to show that the definition is is a correct recursive definition, we need to show that both $[U_{\mathcal{B}}\pi]\phi$ and $[\pi]([U_{\mathcal{B}}\pi]\phi)^*$ are of lower complexity than $[U_{\mathcal{B}}\pi]\Box_a\phi$. This clearly holds for the first sentence. For the second sentence, we use the induction hypothesis from (2), and conclude that $([U_{\mathcal{B}}\pi]\phi)^*$ is a sentence of classical modal logic. That means that $([U_{\mathcal{B}}\pi]\phi)^*$ does not contain any program modalities, and therefore $\mathsf{pc}(([U_{\mathcal{B}}\pi]\phi)^*) = 0$. But then $\mathsf{pc}([\pi]([U_{\mathcal{B}}\pi]\phi)^*) = \mathsf{pc}(\pi) < \mathsf{pc}(U_{\mathcal{B}}\pi) \leq \mathsf{pc}([U_{\mathcal{B}}\pi]\Box_a\phi)$. In other words, $[\pi]([U_{\mathcal{B}}\pi]\phi)^*$ is of lower complexity than $[U_{\mathcal{B}}\pi]\Box_a\phi$. Using the induction hypothesis, we can immediately conclude that neither $([U_{\mathcal{B}}\pi]\Box_a\phi)^*$ contains any programs.

To see that $([U_{\mathcal{B}}\pi]\Box_a\phi)^*$ is equivalent to $[U_{\mathcal{B}}\pi]\Box_a\phi$, if $a \in \mathcal{B}$, we use the Ramsey axiom, that says that $[U_{\mathcal{B}}\pi]\Box_a\phi$ is equivalent to $\Box_a[\pi][U_{\mathcal{B}}\pi]\psi$. By induction hypothesis, we know that $[U_{\mathcal{B}}\pi]\phi$ is equivalent to $([U_{\mathcal{B}}\pi]\phi)^*$, and using the induction hypothesis a second time, we know that $([\pi]([U\mathcal{B}\pi]\psi)^*)^*$ is equivalent to $[\pi]([U_{\mathcal{B}}\pi]\psi)^*$. We can then substitute equivalent sentences and conclude that $([U_{\mathcal{B}}\pi]\Box_a\phi)^*$ is equivalent to $[U_{\mathcal{B}}\pi]\Box_a\phi$. □

We have proven that the axiom systems are sound and complete by showing how DEL can be translated into classical modal logic. Independently, Baltag et al. (to appear) found a very similar translation result. An immediate corollary of this proof is that DEL has the same expressive power as classical modal logic.

Corollary 4.14 The language of DEL has the same expressive power as classical modal logic. □

It is interesting to note that a similar fact does not hold if we extend the language of DEL with operators $C_{\mathcal{B}}$ for common knowledge. Baltag et al. (to appear) show that this language has strictly more expressive power than the language $\mathcal{L}^C$ of classical modal logic with common knowledge operators.

I will now turn to a discussion of the sentences provable in the system.

Necessitation and Updates that Preserve Knowledge

One of the noteworthy aspects of the axiom system here, and main difference with the version presented in Gerbrandy (1997c) is that the axiom system does not contain a necessitation rule for the program modalities. Such a rule would be of the form "If $\vdash \phi$, then $\vdash [\pi]\phi$." From the fact that the axiom system is complete, it follows immediately that this rule is a derived rule in DEL, in the sense that each instance of it is a valid proof in DEL.

Proposition 4.15 It holds that

$$\vdash_{\mathsf{DELK}} \phi \text{ implies } \vdash_{\mathsf{DELK}} [\pi]\phi$$

is a derived rule in DEL. □

There is a good reason to omit the rule from the axiom system, however. The reason is that the rule is *not* a derived rule in DELK45 or DELS5, because the necessitation rule is not sound in the class of introspective (and factual) possibilities. The reason is that an update of an introspective possibility may result in a non-introspective possibility, so even if a sentence ϕ is true in all introspective possibilities, this does not mean that $[\pi]\phi$ is true in all introspective possibilities as well. To see this, consider the following example. Consider a simple language with just a single agent, and one propositional variable. Let w and v be possibilities such that $w(p) = 1$, $v(p) = 0$, and $w(a) = v(a) = \{w, v\}$. A picture of w in the form of an unraveled tree looks like figure 4.9:

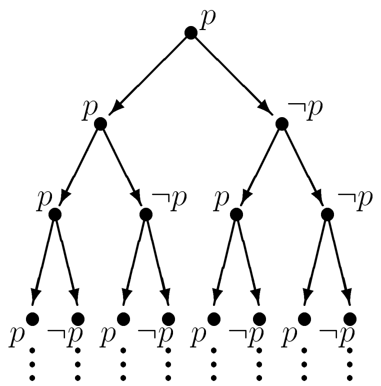


Figure 4.9.

Consider now the update of w with $U_a U_a U_a ?p$. The resulting possibility is neither positively nor negatively introspective, nor is it factual, as can be seen by inspecting figure 4.10.

The fact that updates do not preserve the properties of knowledge or belief is at odds with the characterization of DEL as an *epistemic* logic. This brings us

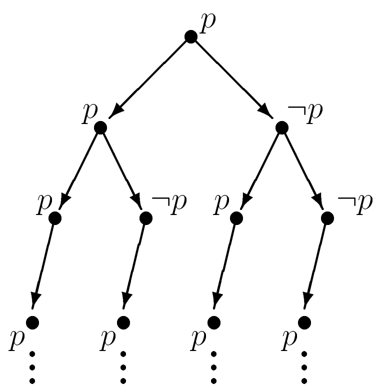


Figure 4.10.

to a choice point in the definition of the semantics. There are two possibilities: we can either redefine the semantics of programs in such a way that programs always preserve the properties of belief, or we can simply leave the semantics of programs as it is.

Most authors that have written about the semantics of information change and its relation with truth have chosen the first option. Baltag et al. (to appear) propose a variation on the semantics of DEL where a program of the form $U_{\mathcal{B}}?\phi$ fails to return an output in case ϕ is false. This would mean that an agent can only learn things that are in fact true, a condition that seems to fit well with our intuitions pertaining to growth of knowledge. Unfortunately (and as the authors note themselves) this condition is not strong enough to guarantee that programs preserve the properties of knowledge either.[3] For suppose that in an S5-possibility where p is true, an agent a learns, consciously, that p is the case The resulting possibility will be one in which the information of a has retained all the properties of knowledge, which is just as we would expect. But the information state of an agent b different from a does not change as the result of a learning new facts. That means that b's information state in the new possibility may not be factive: the world has changed (a has learned something new) without him being aware of it. So, learning things which are true does not guarantee that the resulting possibility will be a representation of knowledge.

Gerbrandy and Groeneveld (1997), and, in a slightly different context, Landman (1986), propose to make only a slight change to the semantics of DEL: declaring updates to be undefined in cases where their output would be a model that does not have the properties of K45 or S5. Clearly, this guarantees that updates

[3] A similar option is taken by Van der Hoek et al. (1994b), who define updates they call 'tests' which roughly correspond to learning whether a certain sentence is true or not. This has as an immediate consequence that agents only learn things that are in fact true. Cf. page 110 for more details about their approach.

always preserve knowledge. A major drawback of this approach is that it changes the logic considerably: sentences which first were valid become invalid in the logic of knowledge, such as $p \wedge \neg[U_a\neg p]\neg p$.

I believe that the second option, not changing the semantics of programs at all, is a better option. First of all, it allows us to choose our axiom system in such a way that the logics of belief and knowledge are conservative extensions of DEL. But more importantly, the semantics of programs given here seems to me basically correct. The observation that updates do not preserve properties such as introspection and reflexivity does not show that the semantics is wrong; it merely shows that the programming language has more expressive power than we perhaps need. If an update with a program results in a state that fails to have certain desired properties, then the fault is not so much with the meaning of the program as in the choice of the modeller to choose the wrong program.

Successful Updates

Very often, if an agent updates with a test, she comes to believe that the test succeeds. We call such updates successful.

Definition 4.16 (successful updates)
An update $U_a?\phi$ is successful in w iff $w \models [U_a?\phi]\Box_a\phi$ □

The update of an information state with a sentence ϕ is successful if ϕ is accepted in the resulting information state. Perhaps surprisingly, updates are not always successful. For example, an update of the information of b with the sentence $\neg\Box_b p \wedge p$ may fail to be successful. Let w be an introspective possibility in which it holds that $w \models p \wedge \neg\Box_b p \wedge \neg\Box_b\neg p$. Consider now the update of the possibility w with the program $U_b?(\neg\Box_b p \wedge p)$, which expresses that b learns p, but she does not believe p. The agent b learns two things from this update: that she does not believe that p (which is true, and, by introspection, is something she knew already), and secondly, that p is true (something she did not know yet). The effect on her information state is that she will eliminate all worlds from her information state where p is false. We can picture this as in figure 4.11. As can be seen, the

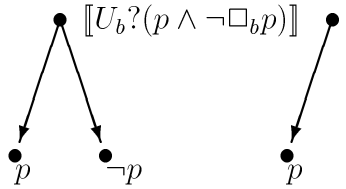


Figure 4.11. An unsuccessful update

update with $U_b?(\neg\Box_b p \wedge p)$ of the possibility pictured by the Kripke model on the left hand side (more precisely, its transitive and euclidean closure) results in the possibility on the right hand side. In this new possibility, the sentence $\neg\Box_b p \wedge p$ is false, while it was true before. Note that in the new possibility, it holds that b knows that $\neg\Box_b p \wedge p$ is false, even though the state is the result after b has made an update with the information that $\neg\Box_b p \wedge p$ is true.

This may be confusing, but it is not paradoxical in any way. The sentence $\neg\Box_b p \wedge p$ is about b's information as well as about the world. If b learns such a sentence, her information state changes: she learns that p from the second conjunct of the sentence. The result is an information state where the first conjunct, which is about the information of b, has become false.

In short, if we analyse 'learning a sentence' with a conscious update, then it makes sense to say that an agent can learn a sentence ϕ, and come to the conclusion that ϕ is false, without deriving a contradiction in the process at all. Later in this dissertation, we will see that the lack of success of updates plays a crucial role in understanding certain puzzles that involve reasoning about knowledge.

The only updates that can fail to be successful are those that express that the agent whose state is updated lacks certain information. In our object language, this means that an update $U_a?\phi$ only fails to be successful if ϕ contains a negative occurrence of the operator $\Box_a$. All other updates are always successful in introspective states. We can generalize this observation to apply to group updates as well.

Proposition 4.17 (successful updates)
If for each $a \in \mathcal{B}$ it holds that all occurrences of $\Box_a$ in ϕ are either positive occurrences, or are in the scope of an operator $\Box_b$ with $b \notin \mathcal{B}$, then:

$$[U_{\mathcal{B}}?\phi]\Box_a\phi \text{ is valid for each } a \in \mathcal{B}.$$

proof: I will use some techniques from section 5.4.1.

Define a $\mathcal{B}$-simulation to be any relation R such that for all w and v, if wRv, then $w[\mathcal{B}]v$ and for each $a \in \mathcal{B}$ and each $v' \in v(a)$ there is a $w' \in w(a)$ such that $w'Rv'$. Write $w \preccurlyeq_{\mathcal{B}} v$ if w and v are related by a $\mathcal{B}$-simulation.

The proposition is a consequence of the following three facts:

(1) If $w[\![U_{\mathcal{B}}?\phi]\!]v$, then there is a $\mathcal{B}$-simulation that connects w with v. This is easily seen by observing that $[\![U_{\mathcal{B}}?\phi]\!]$ is a $\mathcal{B}$-simulation.

(2) If $w \preccurlyeq_{\mathcal{B}} v$, and for each $a \in \mathcal{B}$, each occurrence of $\Box_a$ in ϕ is in the scope of an occurrence of $\Box_b$, with $b \notin \mathcal{B}$, then $w \models \phi$ iff $v \models \phi$. This follows straightforwardly from the fact that w and v differ only in the information states they assign to agents in $\mathcal{B}$.

(3) If $w \preccurlyeq_{\mathcal{B}} v$, and for each $a \in \mathcal{B}$, each occurrence of $\Box_a$ in ϕ is either positive or in the scope of some $\Box_b$ with $b \notin \mathcal{B}$, then $w \models \phi$ implies that $v \models \phi$. This can be seen by observing that ϕ can be rewritten in such a way (using $\wedge$ and $\vee$)

that each positive occurrence of $\Box_a$ does not occur in the scope of a negation at all. The rest is just a simple induction on the complexity of the normal form of ϕ, using (2). □

This proposition can be seen as an observation that 'most' updates are successful.

We will encounter more examples of unsuccessful mutual updates in chapter 6.5, where the puzzle of the dirty children, and the surprise examination paradox are discussed. I will argue there that the lack of success of certain updates is essential for a proper understanding of these puzzles.

The Ramsey Axiom

The so-called Ramsey test says that it is warranted to believe an implication of the form 'If ϕ then ψ' just in case it holds that if you learn that ϕ is the case, then you believe that ψ. Formulated in our framework, the Ramsey test can be formulated as:

$$[U_a?\phi]\Box_a\psi \leftrightarrow \Box_a(\phi \to \psi)$$

This, however, is *not* a valid axiom. The example we used to show that some updates are not successful is a counterexample to the Ramsey test. We saw that there are possibilities in which the sentence $[U_a?\phi]\Box_a\phi$ is false. But clearly, the sentence $\Box_a(\phi \to \phi)$ is a tautology for each ϕ.

For a counterexample to the other direction of the axiom, consider the same possibility as in the previous example, where a does not know whether p is true. The sentence $\Box_a(p \to \Box_a p)$ is false, because a knows that she does not know that p is the case, whereas the sentence $[U_a?p]\Box_a p$ is true.

The failure of the Ramsey test does not mean that there is no connection at all between the conditionals that an agent believes to be true and the effect of the corresponding updates on his information state. There is. Consider axiom 7, which is dubbed the 'Ramsey Axiom.' Using axiom 4, we can immediately derive that the following sentence must be valid:

$$[U_a?\phi]\Box_a\psi \leftrightarrow \Box_a(\phi \to [U_a?\phi]\psi)$$

This is close to the Ramsey axiom, but not precisely the same. What the axiom says is that if a knows ψ after learning ϕ, then a must have already known that if ϕ were true, then ψ would be true after learning ϕ.

The differences between the 'pure' Ramsey test and the version of axiom 7 are relevant only when the consequent of the test, ψ, depends on the information of a itself. If ψ does not contain operators with the agent a in their subscript, the two versions are equivalent.

4.5 Conclusions and Further Work

In this chapter, I have developed a dynamic epistemic logic: a logic for reasoning about information and growth of information in a multi-agent system. First, I established some general desiderata for updates in a multi-agent system: that updates should be *private*, in the sense if the information of an agent changes, the information of others is not affected, and that updates should be *conscious*, in the sense that if an agent learns something, he also learns that he has learned this. It turned out to be relatively easy to define updates with these properties using corecursion in a semantics based on possibilities. It was also straightforward to extend the definition to apply to group updates that affect that common knowledge in a multi-agent system.

Then, I used conscious updates as the basis of the program repertoire of dynamic epistemic semantics, and gave a sound and complete axiom system for the set of sentences that are valid in all models, respectively all belief and all knowledge models. The logic is a conservative extension of static modal logic. In fact, it does not add any expressive power to the classical language at all.

I concluded the chapter by noting some non-standard facts about the logic. One noteworthy fact of DEL is that updates are not successful in general: updating a state with a sentence ϕ does not always result in a state in which ϕ is accepted.

Many open question remain. I'll only mention a few of the more formal open problems.

- Completeness of extensions of DEL with operators of common knowledge and distributed knowledge of chapter 3.

  Baltag et al. (to appear) give a partial answer; they have an axiomatization and a completeness proof for a language with common knowledge operators, but which does not have the full program repertoire.

- The relation algebra of programs.

  The axiomatization of DEL that I have given in this chapter is an axiomatization of the validity of sentences. I have not given an axiomatization of the equalities between programs. Such a logic would have axioms such as:

  $$(U_a?\phi; U_a?p) = (U_a(?\phi; ?p))$$

  But not:

  $$(U_a?\phi; U_a?\psi) = (U_a(?\phi; ?\psi))$$

- The completeness proof I have given in this chapter uses the fact that DEL can be translated into a logic for which we already had a sound and complete axiomatization. In Gerbrandy (1997c), I have given a completeness proof for DEL that is more 'direct.' This proof uses the necessitation rule for programs, and therefore does not work for DELK45 and DELS5. Is there a Henkin-style completeness proof for these logics?

In the next chapter, I will discuss the relation between the semantics of this chapter and other proposals from the literature. In chapter 6.5 I will apply the logic to some puzzles from the literature.

5

Alternatives and Comparisons

This chapter serves two purposes. One goal of this chapter is to compare the semantics of the previous chapter with other approaches to multi-agent information change. There have been several proposals for logics of information change in a multi-agent system. Some authors have approached the topic in a similar way as I did in the previous chapter: by defining operations on model-theoretic structures that correspond to information change. I will discuss, in different degrees of detail, the work of Groeneveld (1995), Landman (1986), Baltag et al. (to appear) and that of Van der Hoek et al. (1994b). In section 5.3 I will discuss the work of Fagin et al. (1995), where information change is considered from quite a different angle. I will show that my work is completely compatible with theirs. Yet another approach is taken by Jaspars (1994), whose work is the topic of section 5.4.

Many of the mentioned authors define information change in terms of operations on Kripke models. This is the second goal of this chapter: to find an interpretation of the programs of DEL in Kripke models that is equivalent (in a sense to be made precise) to the semantics in terms of the non-well-founded possibilities that were used in the previous chapter. One reason for providing DEL with a semantics that is based on Kripke models instead of the possibilities is that the former way of modeling knowledge is much wider known. The hope is that this chapter makes DEL also accessible for researchers that do not want to be bothered with learning a new way of doing set-theory. In particular the results of section 5.2 can be useful in this respect.

5.1 Introduction

If we want to 'mimick' DEL-updates by relations over Kripke models, we need to know when we have done this in a satisfactory way. The following definition is meant to capture this:

Definition 5.1 (soundness and completeness)
Let $\mathcal{K}$ be a class of Kripke models such that each possibility has a picture in $\mathcal{K}$ (in the following, we will only consider interpretations on classes which have this property). Let sol be the function that assigns to each Kripke model in $\mathcal{K}$ the possibility that is its solution. Let $[\![\cdot]\!]^{\mathcal{K}}$ be a function that assigns a relation on $\mathcal{K}$ to each program π. We define:

1. $[\![\cdot]\!]^{\mathcal{K}}$ is *sound* for π iff
   if $(K,x)[\![\pi]\!]^{\mathcal{K}}(K',x')$ then $\mathsf{sol}(K,x)[\![\pi]\!]\mathsf{sol}(K',x')$.
2. $[\![\cdot]\!]^{\mathcal{K}}$ is *complete* for π iff
   if $w[\![\pi]\!]v$ and (K,x) is an element of $\mathcal{K}$ that is a picture of w, then there is a model (K',x') in $\mathcal{K}$ such that (K',x') is a picture of v and $(K,x)[\![\pi]\!]^{\mathcal{K}}(K',x')$.

We say that $[\![\cdot]\!]^{\mathcal{K}}$ is sound and complete for a set of programs iff it is sound and complete for each program of that set. □

In this chapter I will give different interpretations of the programs of DEL that are sound and complete for a certain set of programs. All these interpretations will be given in the following format. First, we define a class $\mathcal{K}$ of Kripke models that contains at least one picture of each possibility. Then we define an interpretation function $[\![\cdot]\!]^{\mathcal{K}}$ that associates with each program a relation on models from $\mathcal{K}$. This function should be sound and complete. Soundness means that no two models stand in the relation $[\![\pi]\!]^{\mathcal{K}}$ if their solutions do not stand in the relation $[\![\pi]\!]$; completeness means that if $[\![\pi]\!]$ has an output on a possibility w, and (K,x) is a picture of w, then $[\![\pi]\!]^{\mathcal{K}}$ has an output on (K,x).

Given such an interpretation function $[\![\cdot]\!]^{\mathcal{K}}$ we can extend the definition of truth of classical modal logic to apply to the whole language of DEL. It is not hard to see that if an interpretation $[\![\cdot]\!]^{\mathcal{K}}$ is sound and complete, it holds for each sentence ϕ of DEL that ϕ is valid iff it is true in all models in the class $\mathcal{K}$ when we interpret programs using the interpretation function $[\![\cdot]\!]^{\mathcal{K}}$. That means that the new interpretation of the language of DEL results in exactly the same logic.

Proposition 5.2 If $[\![\cdot]\!]^{\mathcal{K}}$ is sound and complete (for a set of programs Π), then it holds for each sentence ϕ (in which only programs from Π occur) that:

$$\phi \text{ is true in all models in } \mathcal{K} \text{ iff } \phi \text{ is true in all possibilities.}$$

proof: First, we show by induction on the complexity of ϕ that for each (K,x) in $\mathcal{K}$ it holds that $(K,x) \models \phi$ iff $\mathsf{sol}(K,x) \models \phi$. The proof is a straightforward extension of the proof of proposition 3.8. The new case is when ϕ is of the form $[\pi]\psi$.

First suppose $\mathsf{sol}(K,x) \not\models [\pi]\psi$. There must be a v such that $\mathsf{sol}(K,x)[\![\pi]\!]v$ and such that $v \not\models \psi$. By completeness v has a picture (K',x') in $\mathcal{K}$ such that that

$(K,x)[\![\pi]\!]^{\mathcal{K}}(K',x')$. By induction hypothesis $(K',x') \not\models \psi$, and we can conclude that $(K,x) \not\models [\pi]\psi$.

For the other direction, assume $(K,x) \not\models [\pi]\psi$. Then there is a (K',x') such that $(K,x)[\![\pi]\!]^{\mathcal{K}}(K',x')$ and $(K',x') \not\models \psi$. By soundness $\mathsf{sol}(K,x)[\![\pi]\!]\mathsf{sol}(K',x')$, and by induction hypothesis $\mathsf{sol}(K',x') \not\models \psi$, so it follows that $\mathsf{sol}(K,x) \not\models [\pi]\psi$.

Since we have chosen $\mathcal{K}$ in such a way that each possibility has a picture in $\mathcal{K}$, we now know that for each possibility we can find a corresponding model in which the same sentences are true. Therefore, validity in $\mathcal{K}$ coincides with validity in the class of all possibilities. □

If we interpret programs by a function that is sound and complete, the resulting logic is just DEL. This does *not* mean, however, that if two programs π and π' get assigned the same relation between possibilities in DEL, also $[\![\pi]\!]^{\mathcal{K}}$ and $[\![\pi']\!]^{\mathcal{K}}$ are the same. In other words, the fact that an interpretation is sound and complete does not mean that programs with the same meaning in DEL also are interpreted as the same relation by $[\![\cdot]\!]^{\mathcal{K}}$.

Defining Updates

One unenlightening way of giving a sound and complete interpretation is by taking the correspondence results of proposition 3.7 and use them to map $[\![\pi]\!]$ to a relation on Kripke models:

Definition 5.3 Let $\mathcal{K}$ be the class of all Kripke models and define the relation $[\![\pi]\!]^{\mathcal{K}}$ between Kripke models by setting for any two (K,x) and (K',x'):

$$(K,x)[\![\pi]\!]^{\mathcal{K}}(K',x') \quad \text{iff} \quad \mathsf{sol}(K,x)[\![\pi]\!]\mathsf{sol}(K',x') \qquad \square$$

This way of defining updates on Kripke models is not very interesting. What we are looking for in this chapter is a way of interpreting programs directly on Kripke models without using the machinery of non-well-founded set-theory.

Giving a recursive definition of program operators different from $U_{\mathcal{B}}$ is easy enough.

Definition 5.4
Let $\mathcal{K}$ be a class of Kripke models such that each possibility has a picture in $\mathcal{K}$. We define an interpretation $[\![\cdot]\!]^{\mathcal{K}}$ on $\mathcal{K}$ for tests, sequencing and choice as follows:

$$\begin{array}{rcl}
(K,x)[\![?\phi]\!]^{\mathcal{K}}(K',x') & \text{iff} & (K,x) = (K',x') \text{ and } (K,x) \models \phi \\
(K,x)[\![\pi;\pi']\!]^{\mathcal{K}}(K',x') & \text{iff} & \text{there is a } (K'',x'') \text{ such that} \\
 & & (K,x)[\![\pi]\!]^{\mathcal{K}}(K'',x'')[\![\pi']\!]^{\mathcal{K}}(K',x') \\
(K,x)[\![\pi \cup \pi']\!]^{\mathcal{K}}(K',x') & \text{iff} & (K,x)[\![\pi]\!]^{\mathcal{K}}(K',x') \text{ or } (K,x)[\![\pi']\!]^{\mathcal{K}}(K',x')
\end{array}$$

Note that in this definition tests are interpreted as subsets of the identity relation. That means that this definition assigns different relations to programs than definition 5.3, where tests of the form $?\phi$ are interpreted as the relation that holds between any two bisimilar Kripke models in which ϕ is true.

In any case, the format seems to be a natural one, and the interpretation is sound and complete for these programs, so I will take it as the basic scheme underlying the definitions of this chapter.

What is missing, of course, is a definition of the effect of the program operators $U_{\mathcal{B}}$. We need to add a clause of the form:

$$(K, x)[\![U_{\mathcal{B}}\phi]\!]^{\mathcal{K}}(K', x') \quad \text{iff} \quad (K', x') = (K, x) + U_{\mathcal{B}}\pi$$

In the following sections I will mostly be concerned with defining the operation $+U_{\mathcal{B}}\pi$ in different ways. To show that the interpretations I will give are sound and complete, I will often use the following lemma:

Lemma 5.5 Suppose $+U_{\mathcal{B}}\pi$ is an operation that is defined for all Kripke models in a class $\mathcal{K}$. If it holds that

$$\mathsf{sol}(K, x)[\![U_{\mathcal{B}}\pi]\!]\mathsf{sol}((K, x) + U_{\mathcal{B}}\pi)$$

Then the interpretation $[\![\cdot]\!]^{\mathcal{K}}$ given by the clauses above is sound and complete.

proof: Soundness can be proven by a simple proof on the structure of the programs; the step for programs of the form $U_{\mathcal{B}}\pi$ is given by assumption. The proof of completeness has the same structure, where the step for programs of the form $U_{\mathcal{B}}\pi$ follows from the assumption that $+U_{\mathcal{B}}\pi$ is total in $\mathcal{K}$ (that is, its value is always defined) together with the assumption that $[\![U_{\mathcal{B}}\pi]\!]^{\mathcal{K}}$ is sound. □

5.2 Updates on Kripke Models

The most natural way of defining updates on Kripke models is by removing, replacing and adding worlds or arrows. I will explore this approach in this section. First, I will discuss a very simple case, that concerns updates with tests in which all agents are involved. It turns out that such updates can be straightforwardly modeled by simply removing arrows from Kripke models. However, this naive method will not work for more complex updates. In the rest of this section I will discuss some more involved operations on Kripke models that provide the tools to give a sound and complete interpretation for all programs of DEL.

5.2.1 Removing Arrows

The effect of an update of the form $U_{\mathcal{A}}?\phi$ is that all agents consciously and mutually learn that ϕ. We can straightforwardly define the effect of a conscious update by all agents with the test that ϕ on a model K by removing from K all edges that lead to worlds where ϕ is false.

Definition 5.6 Let $K = (W, (\stackrel{a}{\longrightarrow})_{a \in \mathcal{A}}, V)$ be a Kripke model. We define $K + U_{\mathcal{A}}?\phi$ to be the model $(W, (\stackrel{a}{\leadsto})_{a \in \mathcal{A}}, V)$, where $\stackrel{a}{\leadsto}$ is given by:

$$x \stackrel{a}{\leadsto} y \text{ iff } x \stackrel{a}{\longrightarrow} y \text{ and } (K, y) \models \phi$$

For pointed Kripke models, we set $(K, x) + U_{\mathcal{A}}?\phi = (K + U_{\mathcal{A}}?\phi, x)$. □

The model $(K, x) + U_{\mathcal{A}}?\phi$ is just as the model (K, x), except that all edges leading to worlds at which ϕ is false (in K) are removed. This definition gives us a semantics that is sound and complete for a fragment of the full programming repertoire.

Proposition 5.7 The interpretation given by definition 5.4 together with definition 5.6 is sound and complete for all programs π for which it holds that if $U_{\mathcal{B}}\pi'$ is a subprogram of π, then π' is a test, and $\mathcal{B}$ is the set of all agents.

proof: Remember that with lemma 5.5 it is enough to show that:

$$\mathsf{sol}(K, x)[\![U_{\mathcal{A}}?\phi]\!]\mathsf{sol}((K, x) + U_{\mathcal{A}}?\phi)$$

I will define a relation R on possibilities that is a bisimulation.

$$\begin{array}{lll} vRu & \text{iff} & \text{there is a model } (K, x) \text{ such that} \\ & & \mathsf{sol}(K, x)[\![U_{\mathcal{A}}?\phi]\!]v \text{ and } u = \mathsf{sol}((K, x) + U_{\mathcal{A}}?\phi) \end{array}$$

With the fact that bisimilar possibilities are equal the wanted result follows from the fact that R is a bisimulation. To see that R is a bisimulation, take any v and u such that vRu. Then there must be a (K, x) as in the definition above. Let $w = \mathsf{sol}(K, x)$, and let K' be the model $K + U_{\mathcal{A}}?\phi$.

Clearly, $w(p) = v(p)$ for each $p \in \mathcal{P}$.

For the first bisimulation clause, take any $v' \in v(a)$. Then there must be a $w' \in w(a)$ such that $w' \models \phi$ and $w'[\![U_{\mathcal{A}}?\phi]\!]v'$. But that means that there is an x' such that $w' = \mathsf{sol}(K, x')$ and $x \stackrel{a}{\longrightarrow} x'$ and $(K, x') \models \phi$. This implies that $x \stackrel{a}{\leadsto} x'$ in K'. And that means that $\mathsf{sol}(K', x') \in u(a)$. Combining all this, we find a possibility $\mathsf{sol}(K', x') \in u(a)$ such that $v' R \mathsf{sol}(K', x')$.

For the other direction, take any $u' \in u(a)$. Then there must be an x' such that $\mathsf{sol}(K', x') = u'$ and $x \stackrel{a}{\leadsto} x'$ in K'. By definition of K' it follows that $x \stackrel{a}{\longrightarrow} x'$ and $(K, x') \models \phi$. We can conclude that $\mathsf{sol}(K, x') \in \mathsf{sol}(K, x)(a)$ and (using the induction hypothesis) that $\mathsf{sol}(K, x) \models \phi$. The possibility v' for which it holds that $\mathsf{sol}(K, x')[\![U_{\mathcal{A}}?\phi]\!]v'$ is an element of $v(a)$. It follows that $v'Ru'$, which finishes the proof. □

So, giving an interpretation in Kripke models of mutual updates with tests that involve all agents is fairly easy: just eliminate all arrows that lead to worlds in which the test fails.

Unfortunately, more complicated updates cannot be modeled by simply eliminating edges. To see this, consider the model of figure 5.1. If we want to model

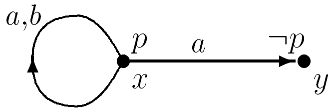


Figure 5.1. A model.

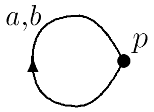


Figure 5.2. No a-arrows to $\neg p$-worlds.

the result of a learning that p is true in x by removing all a-arrows leading to worlds in which p is false, we are left with the model of figure 5.2. In the new model of figure 5.2, a knows that p is the case. But the result is not quite right. The point is that also the information state of b has changed. In the new model $\Box_b\Box_a p$ is true, while previously this was not the case.

The problem with modeling updates by removing arrows (or worlds) from a Kripke model is that arrows (and worlds) can play several roles in a Kripke model. For example, the horizontal a-arrow that connects x with y in figure 5.1 is a witness of at least two logically independent facts that are true in x: that a does not believe that p, but also that b believes that a does not believe that p. If we remove the arrow (or the world y) from the model to model change in the information of a in x, also the information of b in x will change.

A semantics of information change that uses the format of definition 5.6 to define updates in which a proper subset of the whole group of agents is involved will have all kinds of unwanted predictions. The problem is that an update of the information of an agent a with the information that p may have unwanted side effects: the information of other agents about the information of b can change, but also the information that a has about the information of other agents may change.

I am stressing this point, because some of the systems of information change that have been proposed suffer from exactly this problem. Landman (1986) defines updates exactly as in definition 5.6, but also applies the definition to updates that do not involve all agents. Landman notes that his definitions suffer from the problem sketched, and discusses it in some detail, but he does not provide a detailed solution. In the article of Van der Hoek et al. (1994b) the notion of a 'test' is introduced as an operation of epistemic change that corresponds to an agent learning whether a certain sentence is true. A similar operation in Van der Hoek et al. (1994a) is called an 'epistemic update.' In Van der Hoek et al. (unpublished) the notion of 'expansion' is motivated along the same lines as an update in DEL. Tests, epistemic updates and expansions are defined as operations that eliminate arrows from the model in a similar way as it is done in definition 5.6. None of the three operations have the effect of eliminating all arrows, and there are subtle differences between them, but in all three definitions the action that a learns that p in the model of figure 5.1 is modeled by a transition to the model of figure 5.2. The authors do not seem to have noted that this is a problem.

5.2.2 Replacing Arrows

There are many Kripke models in which situations such as given above do not occur: those models in which there is only one 'path' of accessibility relations between any two worlds. Such models can be drawn in the form of a tree.

Definition 5.8 A Kripke model K is a tree iff it holds that if $x \xrightarrow{a_1} x_1 \ldots \xrightarrow{a_n} x_n$ and $x \xrightarrow{b_1} y_1 \ldots \xrightarrow{b_n} y_m$ are paths in K and if $x_n = y_m$, then $m = n$ and $x_i = y_i$ for each $i < n$. □

I will use the symbol $\mathcal{T}$ to denote the class of all trees, and use T as a meta-variable for elements of $\mathcal{T}$.

Trees seem to be the kind of Kripke models that are suitable to define operations of information change on, because the accessibility relations of the different agents interfere in a minimal way in a tree.

In chapter 4 I have informally described he effect of a conscious update expressed by programs of the form $U_{\mathcal{B}}?\phi$ as "everybody in the group $\mathcal{B}$ learns that ϕ, everyone learns that everyone learns that ϕ, everyone learns that everyone learns that ...., etc, *ad infinitum*." This suggests that we can see the effect of a conscious update as the limit of a construction. It is this idea which we will exploit now. Slightly less informal, a mutual update in a group $\mathcal{B}$ with a program π will be defined as the limit of the following construction:

1. First replace all worlds that can be reached from the top node by an edge labeled with an agent in $\mathcal{B}$ by their updates with π.

2. In the new model, replace all worlds that can be reached from the top node by a $\mathcal{B}$-path[1] of length 2 by their updated versions.

$n+1$. Take the model constructed at stage n, and replace all worlds that can be reached from the top node by $\mathcal{B}$-paths of length $n+1$ by their updated versions.

[1] A $\mathcal{B}$-path is a sequence $t_0 \xrightarrow{a_1} t_1 \ldots \xrightarrow{a_n} t_n$ with $a_i \in \mathcal{B}$ for each i. Its length is n.

What I will do is define the separate stages first, and then show how the wanted result is, in a certain precise sense, the limit of this construction. Before doing this, we first need the notion of substituting a set of trees for a subtree of a certain model; this we will use to model the effect of applying a (non-deterministic) program to a subtree of T.

Definition 5.9 (substitution in tree models)
Let (T, t_0) be a tree, and suppose t is not the root of T, i.e. that $t_0 \neq t$. Let $\mathcal{T}$ be a set of trees.

The tree $(T, t_0)[t/\mathcal{T}]$, the model obtained from (T, t_0) by substituting the trees in $\mathcal{T}$ for t, has as its domain the disjoint union of all domains of models in $\mathcal{T}$ and the domain of T. The accessibility relations in the new model consist of the union of the accessibility relations of the old models, with one difference, we remove the edge leading to t, and let $x \xrightarrow{a} t'$ in the new model iff (T', t') is a tree from $\mathcal{T}$, and $x \xrightarrow{a} t$ in T. □

I have tried to visualize this operation in the following figures. Figure 5.3 represents a tree in which the subtree that will be replaced is marked. Figure 5.4 shows the tree that results when we replace the marked subtree by the set $T_1, T_2 \ldots$ of trees.

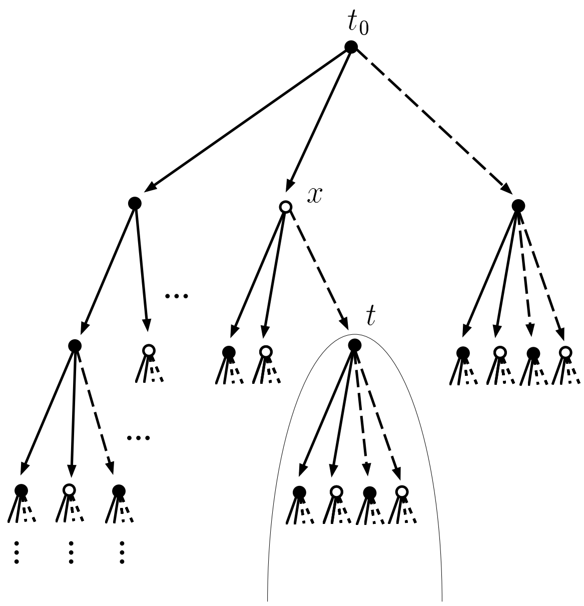


Figure 5.3. A tree.

I will use the substitution operation to define a conscious update on trees. The idea is that given an interpretation of π as a relation over trees, we can model the effect of an update with $U_{\mathcal{B}}\pi$ by recursively replacing the appropriate successors

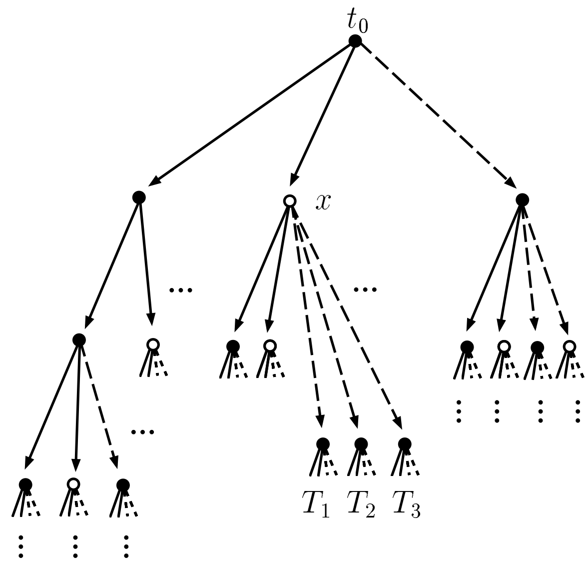


Figure 5.4. The tree of figure 5.3 with t replaced by $\{T_1, T_2, T_3\}$

of the topnode of the tree by the set of trees that are possible outputs of the program π on the subtree.

Definition 5.10 Let (T, t_0) be a tree (with t_0 as its root) and let an interpretation of $[\![\pi]\!]^{\mathcal{T}}$ as a relation between tree models be given.

We define, by induction on n, the models T_n as follows:

- $T_0 = T$
- T_{n+1} is like T_n, except that if $t_0 \xrightarrow{a_1} t_1 \ldots \xrightarrow{a_{n+1}} t_{n+1}$ is a path in T_n, with $a_i \in \mathcal{B}$ for each $i \leq n+1$, then we substitute the set $\{(T', t') \mid (T, t_n)[\![\pi]\!]^{\mathcal{T}}(T', t')\}$ for the node t_{n+1}.
- We define the limit T_ω of this construction in the obvious way: from each T_n in the construction, we take the paths of length n that start with t_0, and let T_ω be the model with exactly those paths. (Note that if there is a path of length n in T_n, it is a path in all models T_m for $m \geq n$ as well.) More precisely, we define $w \longrightarrow^{\omega} v$ iff there is a path $t_0 \longrightarrow^{n} \ldots t_{n-1} \longrightarrow^{n} t_n$ in T_n such that $w = t_{n-1}$ and $v = t_n$.

Finally, we set $(T, t_0) + U_{\mathcal{B}}\pi$ to be (T_ω, t_0). $\square$

We can now use the operation $+U_{\mathcal{B}}\pi$ and 'plug' it into the definition 5.4. We then get a function $[\![\cdot]\!]^{\mathcal{T}}$ that assigns to each program of DEL a relation over trees. The following proposition shows that this interpretation is indeed of the right kind:

Proposition 5.11 The interpretation $[\![\cdot]\!]^{\mathcal{T}}$ is sound and complete with respect to $[\![\cdot]\!]$.

proof: The proof is very similar to that of proposition 5.7. Correctness is the hard part, completeness follows easily. The proofs are by induction on the structure of programs, where the only interesting case is $U_{\mathcal{B}}\pi$. To show that $[\![\cdot]\!]^{\mathcal{T}}$ is correct, we need to show that:

$$\mathsf{sol}(T,t_0)[\![U_{\mathcal{B}}\pi]\!]\mathsf{sol}((T,t_0)+U_{\mathcal{B}}\pi).$$

This follows from the fact that the following relation is a bisimulation:

$$\begin{array}{lll} vRu & \text{iff} & \text{either } v=u\text{, or there is a } (T,t_0) \text{ such that} \\ & & \mathsf{sol}(T,t_0)[\![U_{\mathcal{B}}\pi]\!]v \text{ and } u=\mathsf{sol}((T,t_0)+U_{\mathcal{B}}\pi). \end{array}$$

To see that R is a bisimulation, take any v and u such that vRu. Clearly, v and u agree on the propositional variables.

Now take any a. Suppose first that $a \notin \mathcal{B}$. By definition of T_ω (and the fact that T is a tree), if $a \notin \mathcal{B}$ and $t_0 \xrightarrow{a} t_1$ in T_ω, then the submodel generated by t_1 in T_ω is just the submodel generated by t_1 in T, the tree we started with. But that means that $\mathsf{sol}(T_\omega,t_0)(a)=\mathsf{sol}(T,t_0)(a)$. Since identity is included in R by definition, it follows trivially that all elements of the first set are connected by R to the elements of the second set and vice versa.

The difficult case is where $a \in \mathcal{B}$.

For one direction of the bisimulation condition, take any $v' \in v(a)$. Then there is a $w' \in \mathsf{sol}(T,t_0)(a)$ such that $w'[\![\pi]\!]w''[\![U_{\mathcal{B}}\pi]\!]v'$. This implies that there must be a t_1 such that $t_0 \xrightarrow{a} t_1$ in T and $\mathsf{sol}(T,t_1)=w'$. By induction hypothesis, there must be a (T',t') such that $(T,t_1)[\![\pi]\!]^{\mathsf{tree}}(T',t')$. That means that it holds that $t_0 \xrightarrow{a} t'$ in the models T_1 and that the model generated by t' in T_1 is isomorphic to (T',t'). But then it follows from the way we defined T_ω that the tree (T'_ω,t') is a subtree of T_ω and that $t_1 \xrightarrow{a}_\omega t'$.

Now let $u'=\mathsf{sol}((T',t')+U_{\mathcal{B}}\pi)$. Combining all of the above, we conclude that $u' \in \mathsf{sol}((T,t_0)+U_{\mathcal{B}}\pi)(a)$ and that $v'Ru'$.

For the other direction, take any $u' \in u(a)$. There must be a t_1 such that $t_0 \xrightarrow{a}_\omega t_1$. That means that there is a t'_1 such that $t_0 \xrightarrow{a}_0 t'_1$, that $(T_0,t'_1)[\![\pi]\!]^{\mathcal{T}}(T',t_1)$, and that $(T',t_1)+U_{\mathcal{B}}\pi=(T'_\omega,t_1)=(T_\omega,t_1)$. By induction hypothesis, this implies that $\mathsf{sol}(T_0,t')[\![\pi]\!]\mathsf{sol}(T',t_1)$, and therefore that the unique v' such that $\mathsf{sol}(T',t_1)[\![U_{\mathcal{B}}\pi]\!]v'$ is an element of $v(a)$. The proof shows that $u'Rv'$, which finishes this part of the proof. □

5.2.3 Adding Arrows

Baltag et al. (to appear) interpret the complete set of programs of DEL directly on Kripke models, and show that this interpretation is, in our terminology, sound and complete with respect to DEL-programs.

I will present their definition of the meaning of programs of the form $U_{\mathcal{B}}?\phi$ in this subsection. For the definition of the full program repertoire I refer to the article cited. At the heart of their interpretation of programs stands the following construction:

Definition 5.12 Let $K = (W, (\xrightarrow{a})_{a\in\mathcal{A}}, V)$ be a Kripke model. If ψ is a sentence, and $\mathcal{B} \subseteq \mathcal{A}$, we define the model $K + U_{\mathcal{B}}?\phi = K'$ as follows:

The domain W' of K' consists of the disjoint sum $W + W$ (i.e. the set $(\{0\} \times W) \cup (\{1\} \times W)$. We indicate the left injection of W into $W + W$ (the function taking x to $\langle 0, x\rangle$) by new, and the right injection by old.

In the new model, the accessibility relations are defined as follows:

$$\begin{array}{ll}
\mathsf{new}(y) \xrightarrow{a} \mathsf{new}(z) & \text{if } y \xrightarrow{a} z,\ a \in \mathcal{B}, \text{ and } z \models \phi \\
\mathsf{new}(y) \xrightarrow{a} \mathsf{old}(z) & \text{if } y \xrightarrow{a} z \text{ and } a \notin \mathcal{B} \\
\mathsf{old}(y) \xrightarrow{a} \mathsf{new}(z) & \text{never holds} \\
\mathsf{old}(y) \xrightarrow{a} \mathsf{old}(z) & \text{if } y \xrightarrow{a} z
\end{array}$$

Finally, we set:

$$V'(p) = \{\mathsf{new}(y), \mathsf{old}(y) \mid y \in V(p)\} \qquad \square$$

The idea behind the construction is as follows. Given a model (K, x), we want to construct a pointed Kripke model $(K + U_{\mathcal{B}}?\phi, y)$ that represents the update of (K, x) with $U_{\mathcal{B}}?\phi$. First we make sure we have 'enough' nodes in our model; as we have seen in the examples in section 4.2, sometimes the original model (K, x) does not contain enough nodes to represent the result of updating the model. We do this by taking for the domain of $K + U_{\mathcal{B}}?\phi$ two copies, $\mathsf{new}(y)$ and $\mathsf{old}(y)$, of each node y in K. The nodes $\mathsf{new}(y)$ in the new model are to represent the result of updating (K, y) with $U_{\mathcal{B}}?\phi$, while the nodes $\mathsf{old}(y)$ are to represent the same situations as (K, y).

Given this, defining the new accessibility relations $\xrightarrow{a}$ and the valuation function is straightforward. For suppose that $y \xrightarrow{a} z$ in the old model. As said, $\mathsf{old}(y)$ is to represent that same situation as y did, so we set $\mathsf{old}(y) \xrightarrow{a} \mathsf{old}(z)$. The world $\mathsf{new}(y)$ represents the world y updated with $U_{\mathcal{B}}?\phi$. That means that if $a \notin \mathcal{B}$ the information state of a should not change in any way, so we set $\mathsf{new}(y) \xrightarrow{a} \mathsf{old}(z)$. Finally, when $a \in \mathcal{B}$, the new state of a will consist of all old worlds where ϕ is true, updated with $U_{\mathcal{B}}?\phi$. So, we set $\mathsf{new}(y) \xrightarrow{a} \mathsf{new}(z)$ iff $(K, z) \models \phi$.

Just as in the previous case, we can prove that the semantics is sound and complete with the following lemma:

Lemma 5.13 $\mathsf{sol}(K, x)[\![U_{\mathcal{B}}\phi]\!]\mathsf{sol}(K + U_{\mathcal{B}}\phi, \mathsf{new}(x))$

proof: The proof is just as before. We show that the following relation is a bisimulation, from which the lemma follows:

$$\begin{array}{lll} vRu & \text{iff} & v = u \text{ or} \\ & & \text{there is a model } (K,x) \text{ such that } \mathsf{sol}(K,x)[\![U_{\mathcal{B}}\phi]\!]v \text{ and} \\ & & u = \mathsf{sol}(K + U_{\mathcal{B}}?\phi, \mathsf{new}(x)) \end{array}$$

We show that R is a bisimulation.

Suppose v and u are such that vRu, and let (K,x) be a model for which it holds that $\mathsf{sol}(K,x)[\![U^*_{\mathcal{B}}\pi]\!]v$ and $u = \mathsf{sol}(K + U_{\mathcal{B}}?\phi, \mathsf{new}(x))$. For easier reading, I'll write K^* for $K + U_{\mathcal{B}}?\phi$.

By definition of of $[\![U_{\mathcal{B}}\pi]\!]$ and of V' in the new model it holds that $v(p) = u(p)$ for each propositional variable p.

For the bisimulation clause, suppose first that $a \notin \mathcal{B}$. It is not very difficult to see that for each y in K it holds that $\mathsf{sol}(K^*, \mathsf{old}(y)) = \mathsf{sol}(K,y)$. Since K^* is defined in such a way that $\mathsf{new}(x) \xrightarrow{a} z$ iff there is a y such that $x \xrightarrow{a} y$ and $z = \mathsf{old}(y)$, it holds that $\mathsf{sol}(K^*, \mathsf{new}(x))(a) = \mathsf{sol}(K,x)(a)$. Given the way DEL is defined, we have that $\mathsf{sol}(K,x)[\![U_{\mathcal{B}}?\phi]\!](a) = \mathsf{sol}(K,x)(a)$. In short, if $a \notin \mathcal{B}$, it holds that $v(a) = u(a)$. Since identity is included in R, the rest is a straightforward.

The case where $a \in \mathcal{B}$ is more difficult. Take any $v' \in v(a)$. Then there is a $w' \in \mathsf{sol}(K,x)(a)$ such that $w' \models \phi$ and $w'[\![U_{\mathcal{B}}?\phi]\!]v'$. That means that there is a x' in K such that $x \xrightarrow{a} x'$ in X, $(K,x') \models \phi$, and $\mathsf{sol}(K,x')[\![U_{\mathcal{B}}\phi]\!]v'$. By construction of K^*, it follows that $\mathsf{new}(x) \xrightarrow{a} \mathsf{new}(x')$. But then $\mathsf{sol}(K^*, \mathsf{new}(x')) \in \mathsf{sol}(K^*, \mathsf{new}(x))(a) = w'(a)$. By definition of R it holds that $v'R\mathsf{sol}(K^*, \mathsf{new}(x'))$.

For the other direction, take any $u' \in u(a)$. Then $u' = \mathsf{sol}(K^*, \mathsf{new}(y))$ for some y such that $\mathsf{new}(x) \xrightarrow{a} \mathsf{new}(y)$. By construction of K^*, that means that $(K,y) \models \phi$. Let v' be the (unique) possibility such that $\mathsf{sol}(K,y)[\![U_{\mathcal{B}}?\phi]\!]v'$. Clearly, it holds that $v'Ru'$. Since $\mathsf{sol}(K,y) \in \mathsf{sol}(K,x)(a)$, we may conclude that $v' \in \mathsf{sol}(X,x)[\![U_{\mathcal{B}}\phi]\!](a)$. □

5.3 Reasoning about Knowledge

In the fourth chapter of the book by Fagin et al. (1995) a general framework for modeling knowledge in dynamic distributed systems is developed. In this section, I will show how a certain subset of the programs of DEL can be reconstructed in this framework. The result is an interpretation of a subset of the programs of DEL on a certain class of Kripke models that is sound and complete in the sense of definition 5.1. I will conclude this section with some general marks about the relation between the approach of Fagin et al. and the work presented in this dissertation.

I will give a very condensed presentation of the ideas of Fagin et al. (1995). For further details and many examples I refer to the book.

Knowledge

The concepts of a *local state* and a *global state* play an essential role in the model of Fagin et al. (1995). A local state is simply a description of an agent: it consists of a number of properties that can be relevant for the epistemic state that is associated with an agent, such as "register x has value 22341" or "the formula $p \to (q \vee r)$ is present in the agent's database" or even "neuron 203521 is in state X34."

Given a set of agents, a global state consists of a local state for each of the agents together with a description of the 'environment' that contains all other relevant information. If we denote the agents by the natural numbers $1 \ldots n$, a global state can be seen as an $n+1$-tuple $s = (s_e, s_1, \ldots s_n)$, where s_e is the state of the environment and $s_1 \ldots s_n$ are description of the local states of the respective agents.

Local states of the agents need not make reference to any intensional notions at all. One of the philosophically more interesting aspects of the approach of Fagin et al. is that it provides us with a mathematically very precise way of moving from a model that makes no explicit reference to intensional notions of the agent to a model in which we can talk about the 'knowledge' of these agents. This is done by constructing a Kripke model on the basis of a given set of global states by taking the global states as the set of possible worlds, and define for each agent an accessibility relation on the basis of their local states.

Definition 5.14 (constructing knowledge models)

- A *global state* for a set of agents $\{1, \ldots, n\}$ is an $n+1$-tuple $(s_e, s_1 \ldots s_n)$
- Let S be a set of global states, and let V be a valuation function that assigns to each propositional variable from a given set $\mathcal{P}$ a subset of the set of global states. The *knowledge model* K^S *derived from* S *and* V is the Kripke model $(S, (\xrightarrow{i})_{1 \leq i \leq n}, V)$, where for any two s and t in S, $s \xrightarrow{i} t$ iff $s_i = t_i$. □

All that an agent is 'aware of' is her own local state, so an agent i cannot distinguish between two global states in which she is in the same local state s_i: the only thing that i knows about the global state is that her local state is s_i. That means that if the local state of an agent i is the same in s and in t, we should consider t to be an epistemic possibility for i in s.

Note that all accessibility relation in K^S will be equivalence relations: K^S is an S5-model.

Let us consider a simple example of a distributed system and a Kripke model derived from it. Suppose there are n agents that live in an environment. We assume that the agents store sentences of some interpreted language in their memory and that the environment can be represented by a model for that language. So, a global state is an $n+1$-tuple $(s_e, s_1 \ldots s_n)$, where each s_i $(1 \leq i \leq n)$ is a set of sentences of some language $\mathcal{L}$ and s_e is a model for $\mathcal{L}$. We say that a global state s is *correct* just in case all sentences in s_i are true in s_e for each agent i.

To construct an actual Kripke model from a set of states, we need to specify a valuation function. We will do this by treating the sentences of $\mathcal{L}$ as propositional variables and defining a valuation function V by setting $\phi \in \mathcal{L}$ to be true at a global state s just in case ϕ is true in s_e.

Even if the representation language $\mathcal{L}$ does not contain any epistemic operators at all, we can construct a model of the knowledge of the agents, of their knowledge about one another's information, and of the common knowledge in the system.

If S is a set of correct global states, in which the sentences in the local states of the agents are all true in the environment, then the model K^S derived from S and V has the property that if a sentence ψ is a logical consequence in $\mathsf{S5}$ of the sentences in the local state s_i (note that we treat the sentences in s_i as propositional variables) then it is true at s in the Kripke model that i knows that ψ:

$$\text{If } \Box_i \bigwedge(s_i) \models_{\mathsf{S5}} \Box_i\psi \text{ then } (K^S, s) \models \Box_i\psi.$$

The converse direction holds only if S is the set of all correct global states.[2]

Belief

One limitation of this method of basing Kripke frames on a set of states is that the model is always reflexive: we only get $\mathsf{S5}$-models with the construction of definition 5.14. There is another way of basing a Kripke frame on a set of states which gives us proper $\mathsf{K45}$-models. The method plays only a very minor role in the book of Fagin et al. (1995) (the main reference is an exercise, number 4.11), and it can only be used with the kind of application where the beliefs of the agents are represented explicitly in their local states (here, in the form of sentences), which is only one of the many cases that the more general method of constructing $\mathsf{S5}$-models Fagin et al. (1995) can handle.

When we constructed a knowledge model in the example given above, it was natural to consider only correct global states: states where the information represented by the agents was in fact true. For modeling belief, we will also allow for global states in the model in which the sentences in the local states of agents are false in the environment; this is exactly what distinguishes belief from knowledge. We can define the new accessibility relation as follows:

$$s \xrightarrow{i} t \text{ iff } s_i = t_i \text{ and all sentences in } t_i \text{ are true in } t_e.$$

[2]Fagin et al. write that "the agents' knowledge is completely determined by their local states," which is slightly misleading. As this example illustrates, an agent's knowledge in a distributed system is determined by its local state *together with the set of possible global states.* Taking different sets of global states as the basis of the Kripke model will in general give rise to different knowledge ascriptions in the resulting model.

The motivation behind this definition of the accessibility relations is this: each agent believes that the sentences stored in her local state are in fact true. So in a global state s an agent i believes that the global state is a state where (1) her local state is s_i and (2) all sentences of s_i are true in the environment. This is what the definition is meant to capture.

Note that given a set of states, an accessibility relation defined in this way is transitive and euclidean, but not necessarily reflexive: it is a $\mathsf{K45}$-frame. If S is a set that contains correct global states only, then the frame defined like this is just like the frame we defined previously.

We have a similar result for belief models as we had for knowledge models: if K^S is the belief model derived from the set of global states defined above, then:

$$\text{If } \Box_i \bigwedge s_i \models_{\mathsf{K45}} \Box_a \phi \text{ then } (K^S, s) \models \Box_i \phi$$

Again, the converse direction holds only if S is the set of all global states.

We can generalize this method. To get a belief model from a set of global states, we need to know when (the information represented by) a local state is correct in the global state. Call a global state 'a-correct' if the local state of a is correct.

Definition 5.15 (constructing belief models)
An *interpreted belief system* $\bar{S}$ for a set of agents $1 \ldots n$ and a set of propositional variables $\mathcal{P}$ contains three elements:

(1) A set of global states $S \subseteq S_e \times S_1 \times \ldots \times S_n$

(2) For each each i $(1 \leq i \leq n)$ a set $T_i \subseteq S$ of i-correct states in S and

(3) a valuation function V that assigns to each propositional variable $p \in \mathcal{P}$ a subset of S_e of states in which p is true.

Given an interpreted belief system $\bar{S}$, the *belief model derived from* $\bar{S}$ is the model $(S, (\xrightarrow{i})_{1 \leq i \leq n}, V')$, where $s \xrightarrow{i} t$ iff $s_i = t_i$, and t is i-correct; and $V'(p) = \{s \mid s_e \in V(p)\}$ □

This definition gives us a way of constructing belief models from a set of global states in much the same way as we constructed knowledge models, except that to construct a belief model from a set of states and a valuation function, we need to know when states are correct as well.

We can represent each euclidean and transitive Kripke model by an interpreted belief system. To see this, let K be any model in which the accessibility relations are transitive and euclidean. We define an interpreted system by taking the set of all tuples $(w, \sigma_1 \ldots \sigma_n)$ as global states, where w is any world in K and for each agent i, σ_i is the set of worlds accessible for i from w. We say that a global state s is i-correct just in case $s_e \in s_i$, and set the valuation function $V(s)(p) = 1$ just in case p was true at s_e in K. It is not difficult to see that K^S

is isomorphic to K. This means that we can find a set of global states S and a notion of correctness for each Kripke model in such a way that the belief model derived from S is isomorphic to the Kripke model we started with. The fact that we can represent each belief model by an interpreted belief system shows that for modelling purposes, the use of interpreted belief systems does not limit our possibilities in any way.

Changing beliefs

We have seen how to construct both K45- and S5-models from descriptions of local states of the agents and their environment. I will now turn to the question of defining a notion of information change in a global state.

Based on the work of Fagin et al. we can give an answer that is surprisingly simple. Given a global state s, we get the global state that results after the agents in some given group $\mathcal{B}$ mutually learn that ϕ by simply 'adding' the program $U_{\mathcal{B}}?\phi$ to the local state of each agent in $\mathcal{B}$ by concatenation. In the state t that results the local state of each agent a in $\mathcal{B}$ will simply be the pair $\langle s_a, U_{\mathcal{B}}?\phi\rangle$. Since the information state of the agents outside of $\mathcal{B}$ does not change, one might suspect that their local state remains the same after a $\mathcal{B}$-update. This, however, does not work. For remember that the information state of the agents is not determined by their local states s_a alone, but by the whole set of (a-correct) global states where a's local state is identical to s_a. By adding new states t in which $t_a = s_a$, we would change the information state of a in s in the derived belief model. To make sure we do not ruin our model by adding such states, we have to make sure that the new state of a is different from the old one, also if $a \notin \mathcal{B}$. To do this, we introduce a 'null-element,' $\diamond$, and define the new state of a to be the old state concatenated with $\diamond$.[3]

As always, I assume that the environment does not change as a result of a belief update, so if t is the result of updating s with $U_{\mathcal{B}}?\phi$, then $t_e = s_e$.

More precisely, given S, we define the set of global update states S^+ as follows:

Definition 5.16 Let s be a global state.

For each program of the form $U_{\mathcal{B}}?\phi$, define $s + U_{\mathcal{B}}?\phi$ to be the state t for which it holds that $t_e = s_e$; $t_i = \langle s_i, U_{\mathcal{B}}?\phi\rangle$ if $i \in \mathcal{B}$, and $t_i = \langle s_i, \diamond\rangle$ if $i \notin \mathcal{B}$.

We define $s + \diamond = t$ by setting $t_e = s_e$, and $t_i = \langle s_i, \diamond\rangle$ for each i.

If S is a set of global states, we let the set S^+ be the smallest set containing S that is closed under these two operations. It will be useful to have each inductive step from the standard construction of S^+ available for later reference:

- $S^0 = S$
- $S^{n+1} = \{s + U_{\mathcal{B}}?\phi \mid s \in S^n\} \cup \{s + \diamond \mid s \in S^n\}$

[3]In essence, I am defining what Fagin et al. (1995) call a 'synchronous system.'

- $S^+ = \bigcup_{n<\omega} S^n$

So, S^+ is the smallest set of states that includes S and is closed under the operations $+\diamond$ and $+U_{\mathcal{B}}?\phi$ for each ϕ and $\mathcal{B}$. □

We will use the set S^+ as the basis of our definition of mutual information change in belief systems. A local state in any global state in S^n is an $n+1$-ary sequence with a local state from S as its first element, followed by n occurrences of sentences of the form $U_{\mathcal{B}}?\phi$ and $\diamond$'s. Note that all the sets S^n are mutually disjoint, and note as well that for each $s \in S^{n+1}$, there is a unique s' such that either $s = s' + \diamond$ or $s = s' + U_{\mathcal{B}}?\phi$.

What exactly the operation $+U_{\mathcal{B}}?\phi$ does for the beliefs of the agent in the new state is not clear from the definition of S^+ alone. We do not know what the agents believe in the new model before we have specified which states are correct for which agents. We also need to extend the valuation function to apply to the new states.

Definition 5.17 Let $\bar{S}$ be an interpreted belief system. We define the extended belief systems $\bar{S}^n$ and $\bar{S}^+$. First we define $\bar{S}^n$ be induction on n. The step were $n = 0$ is assumed to be given. For $n+1$, we define:

1. The set of global states of $\bar{S}^{n+1}$ is S^{n+1}.

2. We define i-correctness for any $s \in S^{n+1}$ as follows. We know that for each $s \in S^{n+1}$ there is a (unique) $s' \in S^n$ such that either $s = s' + U_{\mathcal{B}}?\phi$ or $s = s' + \diamond$.

   If $s = s' + \diamond$, then s is i-correct just in case s' is.

   If $s = s' + U_{\mathcal{B}}?\phi$, then we say that s is i-correct only if (1) $i \in \mathcal{B}$ (2) s' was already i-correct and (3) $(K^{\bar{S}^n}, s') \models \phi$ (where $K^{\bar{S}^n}$ is the belief model derived from $\bar{S}^n$)

3. We extend V to a valuation function V^{n+1} by setting $V^{n+1}(p) = \{s' \in S^n \mid \exists s \in V(p) \text{ such that } s_e = s'_e\}$. This reflects the assumption that the truth of propositional variables depends on the environment only, and not on the information that the agents have about the environment.

Finally, $\bar{S}^+$ has S^+ as its global states, a state $s \in S^+$ is i-correct iff there is an n such that $s \in S^n$ and s is i-correct in $\bar{S}^n$, and $V^+(s) = V^n(s)$. □

Given a state $s \in S^{n+1}$, a local state s_i is either of the form $\langle s'_i, U_{\mathcal{B}}?\phi\rangle$ or of the form $\langle s'_i, \diamond\rangle$. In the first case, s is meant to represent the local state of an agent who has, mutually with other agent in $\mathcal{B}$, performed an update with $?\phi$. It holds by definition of S^+ that $i \in \mathcal{B}$ and that there is a unique s' such that $s = s' + U_{\mathcal{B}}?\phi$. The state $s' + U_{\mathcal{B}}?\phi$ is correct from the viewpoint of i if s' was

i-correct in the first place, if what she believed before was true already, and if the new sentence ϕ that she has just learned was true as well. In short, the new information state of i is obtained by taking all possibilities from her old state where ϕ is true, and updating them with $U_{\mathcal{B}}?\phi$.

In the second case, the local state of i is of the form $\langle s'_i, \diamond\rangle$. This means that a has not learned anything new. In particular, that means that a believes that no changes have occurred. So s is i-correct just in case s is of the form $s' + \diamond$, i.e. when s is what results from state s' "after nothing has happened."

The behavior of the operator $+U_{\mathcal{B}}$ on global states is closely related to that of the program $U_{\mathcal{B}}?\phi$ on possibilities in DEL. It is a sound and complete interpretation of such programs of DEL.

Proposition 5.18 Let $\bar{S}$ be an interpreted belief system such that each possibility has a picture in $K^{\bar{S}}$. The operation on Kripke models that corresponds to the operation $+U_{\mathcal{B}}?\phi$ on the states of $\bar{S}^+$ is a sound and complete interpretation of the corresponding programs of DEL.

proof: As before, we need to show that it holds for each global state $s \in S^+$ that:

$$\mathsf{sol}(K^{\bar{S}^+}, s)[\![U_{\mathcal{B}}?\phi]\!]\mathsf{sol}(K^{\bar{S}^+}, s + U_{\mathcal{B}}?\phi)$$

The method of proof is familiar from the previous sections. I show that the following relation is a bisimulation:

$$\begin{array}{lll} vRu & \text{iff} & \text{there is a witness } s \in S^+ \text{ such that} \\ & & \mathsf{sol}(K^{\bar{S}^+}, s)[\![U_{\mathcal{A}}?\phi]\!]v \text{ and } u = \mathsf{sol}(K^{\bar{S}^+}, s + U_{\mathcal{B}}?\phi) \end{array}$$

To see that R is a bisimulation, let v and u be possibilities such that vRu and let s be a witness for the bisimilarity of v and u, as in the definition of R. I will write s for the model $(K^{\bar{S}^+}, s)$ and I will use s^+ to denote the model $(K^{\bar{S}^+}, s + U_{\mathcal{B}}?\phi)$.

Since $V(s) = V(s^+)$ by definition of $\bar{S}^+$, it holds that $\mathsf{sol}(s)(p) = \mathsf{sol}(s^+)(p)$ for each p.

Assume now that a is an agent such that $a \notin \mathcal{B}$. This means that $s^+_i = \langle s_i, \diamond\rangle$. I will show that $v(a) = u(a)$. To see this, note first that $v(a) = \mathsf{sol}(s)(a)$ by definition of $[\![U_{\mathcal{B}}?\phi]\!]$. So we only need to show that $\mathsf{sol}(s)(a) = \mathsf{sol}(s^+)(a)$. The argument runs as follows:

$w \in \mathsf{sol}(s)(a)$ iff
there is an a-correct t such that $\mathsf{sol}(t) = w$ and $s_a = t_a$ iff (with lemma 5.19)
there is an a-correct $t + \diamond$ such that $\mathsf{sol}(t + \diamond) = w$ and $s_a = t_a$ iff
there is an a-correct t' such that $\mathsf{sol}(t') = w$ and $s^+_a = t'_a$ iff
$w \in \mathsf{sol}(s^+)(a)$.

Finally, take any $a \in \mathcal{B}$.

Assume first that $v' \in v(a)$. Then there is a t such that $s \xrightarrow{a} t$ and $\mathsf{sol}(t) \models \phi$ and $\mathsf{sol}(t)[\![U_{\mathcal{B}}?\phi]\!]v'$. Since $s \xrightarrow{a} t$, we know that t is a-correct and that $t_a = s_a$.

Consider now the global state $t + U_{\mathcal{B}}?\phi$. Since t is a-correct, $a \in \mathcal{B}$ and $t \models \phi$, it holds that $t + U_{\mathcal{B}}?\phi$ is a-correct. Since $t_a = s_a$, it holds that $(t + U_{\mathcal{B}}?\phi)_a = (s^+)_a$. But then $s^+ \xrightarrow{a} t + U_{\mathcal{B}}?\phi$, and therefore $\mathsf{sol}(t + U_{\mathcal{B}}?\phi) \in \mathsf{sol}(s^+)(a)$. Putting all of the above together we conclude that v' and $\mathsf{sol}(t + U_{\mathcal{B}}?\phi)$ stand in the relation R.

For the other direction, let $u' \in u(a)$. Since $u = \mathsf{sol}(s^+)$, there must be a t' such that $s^+ \xrightarrow{a} t'$ and $\mathsf{sol}(t') = u'$. This implies that t'_a is of the form $\langle t_a, U_{\mathcal{B}}?\phi\rangle$ and that t' is a-correct, and so there must be a t for which it holds that $t + U_{\mathcal{B}}?\phi = t'$, that $t \models \phi$ and that t is a-correct. But then $\mathsf{sol}(t) \in \mathsf{sol}(s)(a)$ and $\mathsf{sol}(t) \models \phi$ and therefore, $\mathsf{sol}(t)[\![U_{\mathcal{B}}?\phi]\!] \in \mathsf{sol}(s)$. □

Lemma 5.19 $\mathsf{sol}(s + \diamond) = \mathsf{sol}(s)$

proof: We show that the relation R that holds between s and $s+\diamond$ is a bisimulation on $K^{\bar{S}^+}$. First note that by definition of $\bar{S}^+$ the values of the propositional variables are the same in s and in $s + \diamond$. Suppose now that $s \xrightarrow{a} t$. Then $s_a = t_a$ and t is a-correct. But then $(s + \diamond)_a = (t + \diamond)_a$ and $t + \diamond$ is a-correct because t is. So $s + \diamond \xrightarrow{a} t + \diamond$.

For the other direction, suppose $s + \diamond \xrightarrow{a} t$. Then $t_a = \langle s_a, \diamond\rangle$ and t_a is a-correct. By definition of a-correctness, if $t_a = \langle s_a, \diamond\rangle$ and t is a-correct, there must be an a-correct s' such that $t = s' + \diamond$. But since $t_a = (s' + \diamond)_a = \langle s_a, \diamond\rangle$ it follows that $s'_a = s_a$. Since s' is a-correct, $s \xrightarrow{a} s'$. □

I have provided a sound and complete interpretation of DEL by defining operations of belief change as operations on interpreted belief systems. In contrast to the work in the previous section, where the operations on trees were closely related to the corresponding operations on possibilities, the model developed in this section is motivated by independent considerations.

If we compare the way information change is modeled in the framework of Fagin et al. with the approach taken by myself, the main difference is probably that in the former model, changes in the beliefs of agents are derived from changes in an underlying structure of global states, while in DEL information change is defined directly as a relation on possibilities themselves. Fagin et al. are interested in knowledge ascriptions to agents whose 'local states' can be described in more or less objective terms, and they are interested in the relationship between changes in the local states of the agents and the influence of these changes on the knowledge of those agents. In contrast, in this dissertation I have tried to develop a model that is independent of how beliefs are actually represented by agents, and studied information change in an abstract way. I like to think of the two approaches as complementary: in this dissertation I have studied a relatively small set of operations of information change in a fairly abstract way, while Fagin et al. provide a general framework for studying specific situations where information changes. If one feels that what is missing in Fagin et al. (1995) is a general account of what it means to learn something, then perhaps this dissertation fills part of the gap.

5.4 Dynamic Modal Logic

In this section we will study a different approach to information change. Based on ideas from De Rijke (1993) and De Rijke (1994), Jaspars (1994) develops a semantics of information change in multi-modal Kripke models. The idea behind the semantics is appealingly simple. One of the desired properties of a definition of information gain is that it reflects a 'minimal change:' if an agent learns that ϕ, then as little as possible should change in the model (apart from the agent learning ϕ). In other words, the effect of an update of the state of an agent a with the information that ϕ is that we move to a new model where a knows that ϕ, but which differs as little as possible from the old model in any other respect. This idea presupposes that we have a way of ordering models according to resemblance. I will argue in this section that in a multi-agent setting, there is no suitable notion of resemblance that we can use to model updates as we have done in DEL.

Let us first be precise about the logic that is proposed in Jaspars (1994).

Definition 5.20 (language of dynamic modal logic)
The language of dynamic modal logic[4] (DML) is similar to that of DEL.

The set of *sentences* of DML is defined, *mutatis mutandis*, as the set of sentences of DEL: it contains a set of propositional variables $\mathcal{P}$ and is closed under conjunction and negation and the operators $\Box_a$ and $[\pi]$ for each agent a and each program π.

The set of *programs* of DML is the smallest set such that for each sentence ϕ of DML, $?\phi$ and μ-$\mathsf{exp}(\phi)$ are programs of DML, and if π and π' are programs of DML, then so are $\pi;\pi'$ and $\pi \cup \pi'$. □

The difference between the language of DEL and DML is that the former contains an operator on programs $U_{\mathcal{B}}$, while the latter has an operator μ-exp that turns sentences into programs. The intended interpretation of μ-$\mathsf{exp}(\phi)$ is something like 'move to a model in which ϕ is true that differs minimally from the model you started from.' To make this notion of 'differing minimally' precise, a possibilities are presumed to be ordered by the amount of information they contain.

More formally, the definitions of tests, sequencing and disjunction are all standard, the interesting case is the definition of μ-exp. This operator is interpreted relative to an information ordering $\leq$.

Definition 5.21 (minimal expansion)

[4] I am presenting only a small fragment of the full dynamic modal logic – the part that I consider relevant for this discussion. The full program repertoire of DML also contains intersection, complement and converse as program operators, plus operators 'con' and 'exp' that change sentences into programs. I have picked only the operator μ-exp, which can be defined in terms of the above. I have made a slight change in the syntax of the language to suit the conventions used in this dissertation.

$$
\begin{array}{lll}
(K,x)[\![\mu\text{-}\mathsf{exp}(\phi)]\!]^{\leq}(K',x') & \text{iff} & (K,x) \leq (K',x') \text{ and } (K',x') \models \phi \text{ and} \\
& & \text{there is no } (K'',x'') \text{ such that} \\
& & (K,x) \leq (K'',x'') < (K',x') \text{ and} \\
& & (K'',x'') \models \phi
\end{array}
\qquad \square
$$

Given an ordering $\leq$ of Kripke models, the program μ-$\mathsf{exp}(\phi)$, when executed in a model (K,x), returns a model (K',x') that is a nearest model that is higher in the information ordering in which ϕ is true. Relevant for Jaspars' work is the idea that we can use this operator to model information change: for example, the effect of a conscious update of an agent a with the information that p can be seen as changing the model as little as possible in such a way that $C_{\{a\}}p$ is true, i.e. that a knows (consciously) that p.

This way of modeling information change is quite elegant, I believe. It is also incomplete. It is essential to know which information ordering to use: the meaning of μ-exp is completely determined by the ordering $\leq$. And how to define a useful ordering on possibilities is far from clear.

5.4.1 More or Less Information.

There are several ways to study the notion of one possibility containing more information than another. For example, given two possibilities we can compare which sentences of epistemic logic are true, and say that one possibility contains more information than another when 'more' sentences are known by the agents. Another approach is to look at the behavior of programs, and say, for example, that one possibility contains more information than another just in case the latter can be reached from the former by an update with a program expressing 'growth of information.' These two approaches have the obvious drawback that their results depend on the semantics and expressive power of the language considered. Instead, I will concentrate on a more general algebraic notion of ordering possibilities.

In Update Semantics (cf. section 4.1) information states are sets of truth assignments to the propositional variables. It is quite straightforward to see when one information state contains more information than another: an agent in state σ has more information than an agent in τ if in σ more possibilities are ruled out than in τ. If we can find for each possibility in τ a possibility in σ where the same propositional variables are true, then τ contains less epistemic alternatives than σ, and therefore contains more information. In other words, the information ordering on states of Update Semantics is the subset ordering.

The subset ordering is useless for comparing information states that represent higher-order information as well, at least not if information is taken to be introspective. The reason is that if w and v are introspective possibilities and $w(a)$ and $v(a)$ are not equal, then $w(a)$ is not a subset of $v(a)$, nor vice versa. What we need is a way of 'lifting' the subset ordering to apply to states that represent higher order information as well.

There are many ways of doing this, but I will discuss only the most straightforward one. The idea is this. When comparing two sets of possibilities σ and τ that represent higher order information as well, we will say that τ contains at least as much information as σ just in case for each possibility v in τ we can find a possibility w in σ such that (1) w and v agree on the values they assign to the propositional variables and (2) for each agent, the state $v(a)$ contains more information than $w(a)$.

Definition 5.22 (information ordering of possibilities)
A *simulation* is a relation R between possibilities such that for each w and v, it holds that if wRv then $w(p) = v(p)$ for each $p \in \mathcal{P}$ and for each $v' \in v(a)$, there is a $w' \in w(a)$ such that $w'Rv'$.

A possibility v contains at least as much information as a possibility w, written as $w \preccurlyeq v$, just in case there is a simulation R such that wRv. □

The definition of a simulation is like a partial definition of a bisimulation. The relation $\preccurlyeq$ is much like the 'hereditary subsituation' relation of Barwise (1989), which is a relation between (non-well-founded) situations that is meant to hold between two situation s_1 and s_2 "if all the information present in s_1 is present in s_2." Also Jaspars (1994) uses simulations over Kripke models as the ordering that underlies his semantics for growth of information (page 99).

If we consider how the ordering compares with the sentences and programs of DEL, we have the following results.

Proposition 5.23

- For each program π, if $w[\![\pi]\!]v$, then $w \preccurlyeq v$.
- If $w \preccurlyeq v$, if ϕ is a sentence in which $\Box$ occurs only positively (not in the scope of an uneven number of negations) and if $w \models \phi$, then $v \models \phi$.

proof: For the first item, show by induction on programs that $[\![\pi]\!]$ is a simulation. For the second item, note that if $\Box$ occurs only positively in a sentence ϕ, then ϕ is equivalent to a sentence in which $\Box$ does not occur in the scope of a negation at all (but it may occur in the scope of a disjunction). The rest of the proof is then a simple induction on the complexity of such sentences. □

The first item states that programs of DEL 'respect the information ordering,' the second item says that the amount of 'what is known' grows along the information ordering, in the sense that if w contains less information than v, then all sentences that are true in w in which $\Box$ occurs only positively (i.e. all sentences that do not state that certain agents do *not* know certain things) are true in v as well.

All these facts suggest that simulations are a good way to capture the notion of one state containing more information than another one. But when we study the properties of the $\preccurlyeq$ ordering in more detail, this is much less obvious. The relation $\preccurlyeq$ is not anti-symmetric, which means that there are possibilities w and v such that $w \preccurlyeq v$ and $v \preccurlyeq w$, but $w \neq v$. This is, I suppose, not according to intuition: if one possibility contains 'at least as much information' as another as well as 'the same or less information,' the two possibilities 'contain the same information,' and one would expect them to be just the same possibility.

Proposition 5.24 The relation $\preccurlyeq$ is not a partial order. It is transitive and reflexive, but not anti-symmetric.

proof: Transitivity of $\preccurlyeq$ can be shown by showing that the composition of two simulations is a simulation as well. Reflexivity is easy to prove by observing that the identity relation is a simulation.

To see that $\preccurlyeq$ is not anti-symmetric, consider figure 5.5.

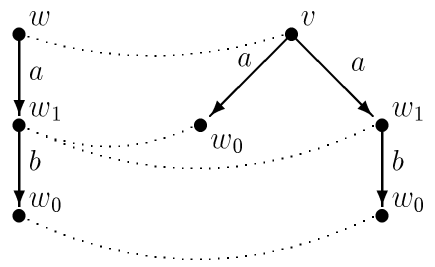


Figure 5.5. $w \preccurlyeq v$ and $w \preccurlyeq v$

It is not hard to see that $w \preccurlyeq v$, since $w(a)$ is a subset of $v(a)$. The dotted lines represent a relation that is a simulation connecting v with w, which shows that $v \preccurlyeq w$. The models in the example can be changed into reflexive and introspective models, by taking the appropriate closure of the a and b edges, but then we have to make sure that the truth-values of the propositional variables in the worlds in the models differ. In particular, it should be possible to distinguish w_1 from w_0; otherwise, the models labeled by w and v would represent the same possibility. □

If we want to use simulation $\preccurlyeq$ as the ordering that underlies the use of DML as a logic for multi-agent information change, the fact that the $\preccurlyeq$ is not anti-symmetric is a serious problem.[5] Most notably, if $\preccurlyeq$ is not anti-symmetric, we cannot guarantee that a model in which ϕ is true does not change as a result of a minimal expansion with ϕ. Intuitively, this does not seem to be right. Jan Jaspars,

[5]The problem is not really the lack of anti-symmetry, but the fact there are models that are 'informationally equivalent' while they clearly represent different situations.

who in his dissertation provides some discussion about information orderings does not seem to have noted this problem.[6]

For example, the principle of privacy, that says that information of agents different from a is not affected by an update of a's information, is not valid. The following 'privacy axiom' is not valid (and is therefore not an axiom in Jaspars 'calculus for constructive communication' $\mathcal{C}_3$).

$$\Box_b\phi \rightarrow [\mu\text{-}\mathsf{exp}(\Box_a\psi)]\Box_b\phi \text{ if } a \neq b$$

In fact, given the minimal conditions that Jaspars puts on his information ordering, it is not even guaranteed that an agent may not 'lose' information she has by an μ-exp-update with a sentence she already knows. The relevant example is that of figure 5.5, where the model with the topnode v is a possible output of the program μ-$\mathsf{exp}(\Box_a\top)$. If this program is to model the effect of 'a getting the information that $\top$ is true', then we would expect no change at all in the model, i.e. that the following principle is valid:

$$\phi \leftrightarrow [\mu\text{-}\mathsf{exp}(\Box_a\top)]\phi$$

If we use the ordering $\preccurlyeq$ as the information ordering, this principle does not hold.

As a side remark, note that the problem of defining an anti-symmetric information ordering on Kripke models is specific to the multi-agent case. It does not occur when we restrict the semantics to $\mathsf{S5}$-models with a single agent: in that class of models, the relation $\preccurlyeq$ of definition 5.22 is anti-symmetric, and corresponds with the subset ordering in Update Semantics.[7] That is why De Rijke (1994) can successfully represent Update Semantics in DML.

The example of figure 5.5 seems to show that $\preccurlyeq$ is too weak as a definition of information ordering. Since $v(a)$ contains more epistemic alternatives than $w(a)$ does — $v(a)$ is a strict superset of $w(a)$ — it would seem that a simply knows more in w than in v. If we look at a typical update that brings us from w to v, it would be the program $U_a(?\top \cup U_b?p)$, which expresses that a learns that either b learned that p, or that nothing happened. In a sense, such a program represents that agent a gets less informed about the world: she learns that the world might have changed in a certain way, but she is not sure whether it actually has. So, perhaps we can solve our problems with Jaspars' semantics by finding a stronger, but still useful, notion of information ordering that circumvents the problems.

[6] Jaspars claims that his information ordering is 'the precise formal description of the construction-elimination dynamics.' Jaspars uses partial models; on the total models we are considering here, this claim boils down to saying that simulations are the precise description of the 'elimination semantics,' which are, roughly, updates with tests in DEL. The fact that $v \preccurlyeq w$ in the example on page 127 makes this claim very dubious.

[7] In section 5.5 I will show how one can represent Update Semantics in DEL. The statement that $\preccurlyeq$ corresponds to the subset ordering can be formally represented, in the terminology of that section, by $\sigma \subseteq \tau$ iff $\sigma^a_w \preccurlyeq \tau^a_w$, which is valid.

I have tried to find such a notion, but cannot think of one that is significantly better than simulation.

The problem is that it is virtually impossible to define an intuitively acceptable notion of information ordering that is anti-symmetric. At least, this is what the following example is meant to show.

The possibilities pictured in figure 5.6 are possibilities for three agents a, b and c, and one propositional variable p. A more precise definition of the possibilities w and w' is this. For any two natural numbers i and j, we define the possibilities w^i_j and v^i_j by the following equations:

$$
\begin{array}{rcllrcll}
w^i_j(p) & = & 1 & & v^i_j(p) & = & 0 & \\
w^i_j(a) & = & \{w^i_j\} & \text{if } i=0 & v^i_j(a) & = & \{v^i_j\} & \text{if } i=0 \\
w^i_j(a) & = & \{w^i_j, v^{j+1}_{i-1}\} & \text{if } i>0 & v^i_j(a) & = & \{v^i_j, w^{j+1}_{i-1}\} & \text{if } i>0 \\
w^i_j(b) & = & \{w^i_j\} & \text{if } j=0 & v^i_j(b) & = & \{v^i_j\} & \text{if } j=0 \\
w^i_j(b) & = & \{w^i_j, w^{j-1}_{i+1}\} & \text{if } j>0 & v^i_j(b) & = & \{v^i_j, v^{j-1}_{i+1}\} & \text{if } j>0 \\
w^i_j(c) & = & \{w^{i'}_j \mid i' < \omega\} & & v^i_j(c) & = & \{v^{i'}_j \mid i' < \omega\} &
\end{array}
$$

We define w'^i_j and v'^i_j in the same way (by replacing each occurrence of w and v in the equation above with w' and v'), except that we set $w'^i_j(c) = \{v'^{i'}_j \mid 1 \leq i' \leq \omega\}$ for each i and j. Note that all possibilities defined here are introspective and reflexive.

Now, let w be the possibility w^1_0 and w' be the possibility w'^1_0. Figure 5.6 is a somewhat schematic picture of the possibilities w and w'. It is not very difficult

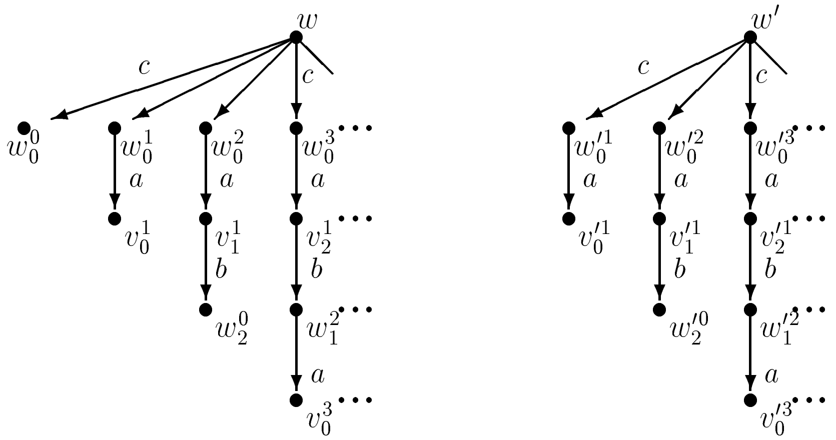


Figure 5.6.

to see that $w \preccurlyeq w'$: the wanted simulation is given by the relation that contains all pairs $\langle w^i_j, w'^i_j\rangle$ and $\langle v^i_j, v'^i_j\rangle$.

Conversely, it holds that $w' \preccurlyeq w$. The reason is that w^0_0 contains more information w'^1_0, w^1_0 contains more information than w^2_0, and, in general, w^i_j contains more information than $w_j'^{i+1}$. So, the relation connecting w_j^{i+1} with $w_j'^i$ and v_j^{i+1} with $v_j'^i$ for each i is the wanted simulation that is a witness for the fact that $w' \preccurlyeq w$.

What we have here is a rather complicated witness of the fact that $\preccurlyeq$ is not anti-symmetric. The main reason for giving this particular example here is that, intuitively, it is not clear at all which of the two possibilities contains more information than the other (the main reason why the example is so complicated is that I wanted to give an example of two S5-models). There are witnesses of the informational equivalence of w and w' in the form of programs in DEL that seem to be typical of 'information growth.'

For one direction, suppose that the agents a, b and c mutually learn in w' that a does not know that p and that b does not know that $\neg p$. Applying this update to w' results in the possibility w. To see this, remember the observation from section 5.2.1 that as an operation on trees such an update with ϕ corresponds to removing all arrows to worlds in which ϕ is false. The only possibilities in the closure of w' where a knows that p are those of the from $w_i'^0$, and the only ones where b knows that $\neg p$ are those of the from $v_j'^0$. Removing the arrows to these worlds in the picture of w' results in a model that is isomorphic to the picture of w. More precisely, it holds in DEL that $w' [\![U_{\{a,b,c\}}(\neg\Box_a p \wedge \neg\Box_a \neg p)]\!] w$. Although this is a complicated update, it is a clear case of a group of agents (mutually) gaining information by eliminating worlds from the model.

For the converse direction, consider the possibility that results if c learns in w_1 that neither a nor b knows that p is the case. This means that she learns that w_0^0 is not an option, and she will therefore eliminate w_0^0 from her information state. If after that, all agents mutually learn that c has learned that neither a nor b knows that p, the resulting state is w'. More precisely, it holds that $w [\![U_c \neg(\Box_a p \wedge \Box_b p); U_{\{a,b,c\}} U_c \neg(\Box_a p \wedge \Box_b p)]\!] w'$. Also this seems to be a clear-cut case of growth of information: c has learned something new by eliminating, consciously, a possibility from her state, en all agents in the group have learned that this happened.

Both updates seem to be cases of proper information growth. A correct notion of information ordering among possibilities should account for this. That means that it cannot be anti-symmetric, and therefore is not useful as the basis for a multi-modal semantics of information change in the style of dynamic modal logic.

To conclude this section: Jan Jaspars' semantics for constructive communication is incomplete at best without a more specific story about what the information ordering is that underlies the use of DML as a logic of multi-agent information change. One of the conditions for such an information ordering as a basis for a workable dynamic epistemic logic (where privacy holds, and where trivial updates do not change a possibility) is that the underlying ordering is anti-symmetric. I have tried to argue that finding a useful ordering that is anti-symmetric is difficult, if not impossible.

5.5 Update Semantics

In Veltman (1996), Update Semantics (cf. section 4.1) is presented as a model of the way a single agent (the hearer) changes her information state as the result of learning sentences. I will show here how Update Semantics can be 'reconstructed' in a multi-agent logic as the logic of conscious updates of a single agent.

We can associate with each introspective possibility w and agent a an information state w^a, which consists of the set of classical worlds (assignments of truth-values to the propositional variables) that correspond to the possibilities in $w(a)$. Vice versa, given a classical world w and an agent a, we can associate with each classical information state s a possibility s^a_w that assigns to each propositional variable the same value as w does and that assigns to a a set containing a possibility s^a_v for each $v \in s$ (a classical information state does not provide us with any information about which agent we are talking about, or what the 'real world' looks like, so we have to supply these parameters ourselves).

Definition 5.25

- If w is a possibility and a an agent, then $w^a = \{v \text{ restricted to } \mathcal{P} \mid v \in w(a)\}$.
- If s is a classical information state, w a classical possible world, and a an agent, then s^a_w is a possibility such that $s^a_w(p) = w(p)$ for each $p \in \mathcal{P}$, and $s^a_w(a) = \{s^a_v \mid v \in s\}$. □

It is not hard to see that s^a_w is an introspective possibility. Using this correspondence between possibilities and US-states as a basis of the comparison, the US-updates can be seen as conscious updates of an introspective information state. More precisely, an update in Update Semantics with a sentence ϕ corresponds to an update in DEL with $U_a?\phi'$, where ϕ' is just as ϕ, except that we replace all occurrences of $\Diamond$ in ϕ with $\Diamond_a$.

Proposition 5.26 For each $\phi \in \mathcal{L}^{\mathsf{US}}$, let ϕ' be just as ϕ but with all occurrences of $\Diamond$ replaced by $\neg\Box_a\neg$. Let $[\cdot]$ be as in definition 4.3. It holds that:

- For all classical information states s and t: $s[\phi]t$ iff $s^a_w[\![U_a?\phi']\!]t^a_w$.
- For all possibilities w and v and each a such that a has introspective information in w: $w[\![U_a?\phi']\!]v$ iff $w^a[\phi]v^a$.
- $\phi_1 \ldots \phi_n \models_{\mathsf{US}} \psi$ iff $\models_{\mathsf{belief}} [U_a?\phi'_1] \ldots [U_a?\phi'_n]\Box_a\psi'$. □

This proposition expresses that if we read formula's of the form "$\Diamond\phi$" as something like "You consider ϕ to be possible," then a US-update can be seen as a conscious update of the information state of an agent who has fully introspective information.

If we combine proposition 5.26 with proposition 4.13, which expresses that every sentence of DEL is equivalent to a classical modal sentence, we have as a corollary that we can reduce the validity of arguments in Update Semantics to validity in classical K45.

Corollary 5.27 Consider the translation function $(\cdot)^*$ of definition 4.11. It holds that:

$$\phi_1 \ldots \phi_n/\psi \text{ is valid in } \mathsf{US} \text{ iff } \vdash_{\mathsf{K45}} ([U_a?\phi'_1] \ldots [U_a?\phi'_n]\Box_a\psi)^* \qquad \Box$$

The observation that Update Semantics can be translated into classical modal logic is not new. Willem Groeneveld (1995) shows how validity in Update Semantics can be reduced to K45-validity, while Eijck and de Vries (1995) show how Update Semantics can be reduced to S5. The present translation adds a new item to this list of translations.

5.6 Conclusions

In this chapter I have compared DEL with some other approaches to modeling information change. I have shown how DEL can be defined by operations on Kripke models. I have given three different ways that may be done. First, I defined updates with tests of sentences ϕ in which the group of all agents is involved simply by removing arrows to each world where ϕ is false. This simple method of modeling information change does not work for updates which do not involve all agents, so in section 5.2.2 I showed how all updates can be defined as operations on trees. In section 5.2.3, I sketched a different approach to modeling information change in Kripke models, one that is taken in Baltag et al. (to appear).

In the section after that, the work of Fagin et al. (1995) was discussed in the light of DEL. We have seen that we can use the methods developed in that book to mimick the effect of programs of DEL in models of knowledge and belief.

In section 5.4, I made a comparison with the work of Jaspars (1994), who uses Dynamic Modal Logic to model information change in a multi-agent system. This approach hinges on the use of an information ordering that can be used to compare different models. We have seen that it is quite difficult to find an information ordering that is intuitively acceptable as well as useful for dynamic semantics. I concluded that for want of a working way of ordering possibilities, this approach cannot be used in a multi-agent setting.

Finally, I have shown how update semantics can be seen as a special case of DEL: the updates of Update Semantics correspond with conscious updates of the information state of an agent with introspective information.

In conclusion, we can say that it is very well possible to interpret the programs of DEL as relations on certain subclasses Kripke models. However, giving such an interpretation is much more cumbersome than defining programs as relations on non-well-founded possibilities.

6

Changing the Common Ground

In the preface to a collection of articles from a workshop on common knowledge, the editors describe the role of the common ground in dialogue as follows:

"Dialogues constitute a field where action and exchange of information meet. [...] Information is incrementally advanced, accepted (or rejected), acknowledged and ratified. This process of so-called grounding may be seen as establishing a common ground of mutually known or believed information presupposed to be available to the participants later on. Grounding has two sides:

- The "external" side concerning the structure and development of the dialogue as a whole [...]

- The "internal" side concerns the more fine-grained aspects of the agent's information processing, including intentions, attitudes and mental states.

It is an important question, whether one of both sides is systematically reducible to the other, and, if so, which is the reduction base."

(Heydrich and Rieser (1998))

In this chapter, I will study and compare simple notions of information change of the common ground from an 'external' and an 'internal' viewpoint. The main result is that even in simple cases such as these, the 'external' viewpoint cannot be reduced to the 'internal' one, nor vice versa. I will try to argue, and, where possible, make precise formally, that under certain minimal assumptions on information change and the way the common ground is represented, the two approaches are incompatible.

6.1 Introduction

The notion of a 'common ground' plays an important role in many models of dialogue. It is a body of public information which changes during the course of a conversation and is used to keep track of what has happened in the conversation, delimits the range of possible further utterances and influences the interpretation of those utterances.

How exactly the common ground should be characterized is not agreed upon. To give some early examples: Lewis (1979) uses the metaphor of a 'conversational scoreboard' on which the relevant information about the 'moves' in the dialogue game are noted. Stalnaker (1978) speaks about 'presuppositions' as that 'what is taken by the speaker to be the common ground of the participants in the conversation, what is treated as their common knowledge or mutual knowledge.' Hamblin (1971) uses the metaphor of a 'commitment slate' that is used to keep track of that information that the participants are committed to on the ground of the utterances they make. Yet other writers simply identify the common ground with 'that what is mutually believed.' Clark and Marshall (1981), for example, argue that it is necessary for a successful use of a definite description that it should be mutual knowledge what the definite refers to.

All of these models have in common the idea that the common ground changes as the result of the actions of agents. These changes can be modeled in (at least) two ways.

In the first kind of model, the 'external' view, a representation of the common ground is taken as primary, and the effect of an utterance in a dialogue is modeled by showing how the utterance affects the common ground.

In the second approach, the 'internal' view, changes in the common ground are derived from changes in the belief states of the agents involved in the dialogue. In this approach one starts with the belief states of the dialogue participants considered separately. The effect of an utterance can then be modeled by the effect it has on the belief states of each of the separate agents. Since the contents of the common ground depend on the belief states of the participants, a change in the states of the participants will lead to a change in the common ground as well.

Consider the following diagram:

$$
\begin{tikzcd}
\text{Common Ground} \arrow[r, "\text{information change}"] & \text{Common Ground} \\
\text{Agents in States} \arrow[u] \arrow[r, "\text{information change}"] & \text{Agents in States} \arrow[u]
\end{tikzcd}
$$

The upper half of the picture represents the external view on changes in the common ground: the changes in the common ground are modeled as an operation from a representation of a common ground to a representation of the new common ground. The upper horizontal arrow represents such an operation of information change.

The diagram as a whole represents the second, 'internal,' view on changes in the common ground. At the bottom corners of the picture, there are agents which have certain information: each of them is in a certain information state. The arrow labeled 'information change' in the bottom part of the picture represents the change in these states that is the effect of an utterance by one of the agents. The vertical arrows represent some way of extracting the common ground from the agents' states.

The leading question in this chapter is whether the internal and the external views on changes in the common ground are compatible. Given the diagram above, the question is whether first taking the common ground in a certain model w of the states of the agents, and then changing the common ground according to some specified way, gives the same result as first changing the model w, and then seeing what the common ground is in the result. More precisely, the question is under which conditions the diagram above commutes.

To make sense of this question, we need to be more precise about filling in the parameters: the kind of representation we use for states of agents and for the common ground, how these two are related, and what information change consists of. Of course, the answer to our question depends for a great deal on how we choose to fill in these parameters.

In the next two sections I will study the diagram using the classical possible worlds framework that I have used throughout this dissertation. First I will discuss the relation between the common ground and common knowledge in more detail in section 6.2. In the section after that I show how a given function that describes information change can be 'lifted' to an operator that models 'mutual information change,' using the techniques discussed in chapters 1 and 4. With this formal machinery, we have the tools to instantiate the informal picture above. We will look at operations of belief change such as expansion, contraction and revision. The main conclusions are negative: the diagram generally does not commute, not even for a relatively simple notion of belief change such as expansion. When considering weaker properties than commutativity, expansion fares fairly well, but I will argue that revision and contraction (and any kind of belief change operation that has certain minimal properties in common with these two) have properties that are incompatible with the assumption that the diagram behaves in a reasonable way.

In section 6.4 I will briefly study the same questions in a more general framework. The results will be similar to those of the preceding sections. The main purpose of this section is to show that the negative results hold for any kind of model that has certain minimal properties in common with the possible worlds approach.

The chapter ends with a section entitled 'conclusions.'

6.2 Mutual Belief and the Common Ground

In section 3.2 I discussed the concept of common knowledge or mutual belief. By definition, a sentence ϕ is mutually believed iff each participant believes that ϕ is true, each participant believes that all other participants believe ϕ is true, etcetera, *ad infinitum*.

As I remarked in the introduction, there are authors that identify the common ground with that which is mutually believed. Such an identification also lies at the basis of the work in this chapter, but it is not altogether uncontroversial to identify the two.

If the common ground is the 'conversational scoreboard,' or a 'commitment slate,' one can argue that the contents of the common ground are independent of what the participants actually believe to be true. With respect to the conversational scoreboard, it does not matter whether a dialogue participant is being honest or telling a lie: the liar is committed to his utterances in the same way as when he would be speaking the truth. Clearly then, there may be sentences on the scoreboard that are not believed, let alone mutually believed.

I think this point is valid, but it does not necessarily imply that the concept of mutual belief is irrelevant to the concept of the common ground. If we want a useful model of the common ground, it should also be useful for conversations in which the participants try *not* to mislead. If we restrict the study of relations between belief change and the common ground to changes in the common ground that arise from honest, unmisleading utterances alone, we will also learn something about the more general case where agents try to mislead.

I will assume in the rest of this chapter that the information in the common ground is in fact believed to be true by each of the participants.

A property of the common ground that, as far as I know, is shared in each model of the common ground that has anything to say about higher-order information (beliefs about beliefs) is that the common ground is in some sense 'publicly accessible:' each of the participants knows what information is in the common ground. Given that whatever is in the common ground is believed by everybody, the public accessibility of the common ground implies that everybody believes that everybody believes the information in the common ground. We can repeat this argument to get arbitrary iterations of 'everybody believes ...'

So, under the assumption that the common ground is mutually accessible, and that all information contained in it is believed, it follows that all information in the common ground is mutually believed.

The answer to the question whether all mutual beliefs are in the common ground depends on the view one takes of that common ground. If the common ground is seen as a kind of conversational scoreboard, or as only containing the information that the dialogue participants are committed to by utterances actually made the answer will be 'no:' surely many facts that are not explicitly stated in the dialogue can be taken to be mutual beliefs, such as the fact that the participants speak a certain language, that the speaker has an enormously big red nose, that there is a vase of flowers on the table between them, etcetera. Other authors, such as Clark and Marshall (1981), argue that all such information should be part of the common ground, because all information that is mutual knowledge can be made use of in a dialogue.

We may conclude that although a complete identification of the common ground with that which is mutually believed is controversial, we can still identify the two in particular idealised cases, and thereby learn something about the general case. This is what I will do in the next sections.

Formal notions of mutual belief

We will represent the common ground in a possibility w by an information state that contains exactly the information that is mutual belief. This information state contains all and only possibilities v for which it holds that one of the agents considers v possible (in w), or that one of the agents considers it possible that one of the agents considers v possible, etcetera. We let the notation $C(w)$ stand for this set of possibilities.

Definition 6.1 (common ground, cf. definition 3.15)
The *common ground* between the agents in a possibility w, $C(w)$, is the smallest $\mathcal{A}$-closed set of possibilities that includes $w(a)$ for each $a \in \mathcal{A}$ □

When we discussed common knowledge, we saw that a sentence is accepted in the state $C(w)$ exactly when it is common knowledge in w:

Fact 6.2 $C(w) \models \phi$ iff $w \models C\phi$ □

Before going back to our diagram I would like to make some remarks about $C(w)$.

First, note that in $C(w)$, we have lost information that was present in w: in general, there are w and v that are different from each other such that $C(w) = C(v)$. This also holds within the class of introspective possibilities. In particular, we cannot see from $C(w)$ alone where its possibilities 'come from:' there is no way of telling from the structure of $C(w)$ whether some $v \in C(w)$ is there because some a thought it possible, or because some a thought some b considered it possible. We will return to this observation later.

Another remark concerns the complexity of $C(w)$: it contains possibilities in which information of agents is represented, the information they have about each other's information, etcetera. Often, in models of dialogue, the common ground is not taken to be that complex at all: sometimes it contains only world-information (information that can be expressed by non-modal sentences), and in general, higher-order information (information about information) is only represented up to some very restricted finite depth. Also this point will be taken up in the next section, where we really start proving things about our diagram.

Zeevat (1997) develops a theory of information change in the common ground that is closely related to the approach adopted in this dissertation, and it may be interesting to make some remarks about its connection with the present work at this point.

Consider the following operation on sets of possibilities that collects all worlds considered possible in one of the possibilities of that set. We call the operation E.[1]

Definition 6.3 $E(\sigma) = \bigcup\{w(a) \mid w \in \sigma, a \in \mathcal{A}\}$ $\Box$

If σ is a singleton set $\{w\}$, we will write $E(w)$ for $E(\{w\})$.

Fact 6.4 $C(w)$ is the smallest set σ containing $E(w)$ such that $E(\sigma) \subseteq \sigma$ $\Box$

In Zeevat (1997) it is argued that a fact is in the common ground just in case it is common ground that everyone knows this fact. More formally, that means that if a set of possibilities σ is a representation of the common ground, it must be a fixed point of E. In other words, it must have the property that:

$$\sigma = E(\sigma)$$

Let's call this property the 'Zeevat property.' It turns out that many, but not all, possibilities have a common ground with the Zeevat property:

Fact 6.5 It holds that $C(w)$ has the Zeevat property iff $E(w) \subseteq E(C(w))$ $\Box$

That $C(w) = E(C(w))$ is not a very strong property of common grounds. For example, it is implied by introspection:

Fact 6.6 If w is introspective, then $E(w) \subseteq E(C(w))$ $\Box$

Each $C(w)$ belonging to an introspective possibility has the Zeevat property. In other words, if we assume that information is introspective, then the view of the common ground as containing exactly all mutual beliefs is perfectly consistent with Zeevat's view on the common ground.

[1] The 'E' is from 'everyone.' Just as $C(w) \models \phi$ iff $w \models C\phi$, so it holds that $E(\{w\}) \models \phi$ iff $w \models \Box_a\phi$ for each a, i.e. just in case 'everyone knows ϕ.'

6.3 Changing the Common Ground

Suppose we are given an operator F over information states that expresses some sort of information change. What I have in mind is an operator such as 'expand with p' or 'revise with ϕ' (Alchourrón et al. (1985)) or the update functions from Update Semantics (Veltman (1996), see also the previous chapters). The first question that I will try to answer here is what it means for a group of agents to apply such a function together; the second question is how such functions behave when we compare the internal with the external approach to changes in the common ground.

Multi-Agent Expansion

A relatively simple kind of operation of information change is that of *expansion*. If we restrict our attention to expanding with sentences that do not contain epistemic operators, then expansion with a certain sentence means simply adding the information contained in that sentence to the information you already have. In the possible worlds framework that is used here, this comes down to discarding all possibilities in which the sentence is false.

Definition 6.7 For each information state σ, $\sigma + \phi = \{v \in \sigma \mid v \models \phi\}$ □

This definition is familiar from Update Semantics. If one takes classical logic as the 'base logic' in the work on belief revision, definition 6.7 is equivalent to the definition of expansion used there.

We are looking for a definition of 'mutual update' on the level of possibilities that corresponds with a change in the common ground. Consider the following definition, in which the notation $+^* \phi$ stands for a mutual expansion with ϕ:

Definition 6.8 (mutual expansion)

$$\begin{array}{lll} w +^* \phi = v & \text{iff} & w[\mathcal{A}]v \text{ and for each } a \in \mathcal{A} \\ & & v(a) = \{u +^* \phi \mid u \in w(a) \text{ and } u \models \phi\} \end{array}$$

Similarly,

$$\sigma +^* \phi = \{w +^* \phi \mid w \in \sigma \text{ and } w \models \phi\}$$ □

In this definition, the notation $w[\mathcal{A}]v$ stands for the fact that w and v differ at most in the states they assign to agents in $\mathcal{A}$: w and v assign the same truth-values to the atomic sentences. I will use this operation to model changes in the common ground, both internally and externally. The operation $+^*$ corresponds with learning that ϕ, and learning that all agents have learned that ϕ.

If we compare this definition with the semantics of DEL, it is easy to see that $w +^* \phi = v$ iff $w[\![U_{\mathcal{A}}?\phi]\!]v$. The idea behind the definition is the same as that behind the definition of a mutual update: one of the participants a learning that all participants have expanded with ϕ is the same thing as a learning ϕ herself, and moreover, changing each of the possibilities in her resulting state to the effect that the participants have mutually learned that ϕ.

One way of viewing the operation $+^*$ is that it models a certain fact becoming common knowledge, as the following fact shows.

Fact 6.9 If ϕ is non-modal, then:

$$w +^* \phi \models C\phi \qquad \square$$

Now that we have given all parts of our diagram a formal interpretation, we can redraw it:

$$\begin{tikzcd}
\sigma \arrow[r, "+^*\phi"] & \tau \\
w \arrow[u, "C"'] \arrow[r, "+^*\phi"] & v \arrow[u, "C"']
\end{tikzcd}$$

This diagram commutes just in case $C(w +^* \phi) = C(w) +^* \phi$. It turns out that this is not the case:

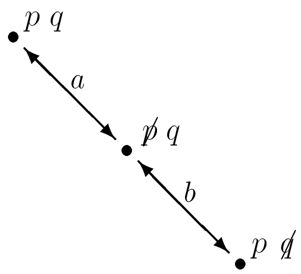
Fact 6.10 There are (introspective and reflexive) w and ϕ such that:

$$C(w +^* \phi) \neq C(w) +^* \phi$$

proof: Consider the three possibilities given by the following equations:

$$\begin{array}{lll}
w_0(p) = 1 & w_1(p) = 0 & w_2(p) = 1 \\
w_0(q) = 1 & w_1(q) = 1 & w_2(q) = 0 \\
w_0(a) = \{w_0, w_1\} & w_1(a) = \{w_0, w_1\} & w_2(a) = \{w_2\} \\
w_0(b) = \{w_0\} & w_1(b) = \{w_1, w_2\} & w_2(b) = \{w_1, w_2\}
\end{array}$$

We can draw this possibility as follows:



In this picture, the topmost dot represents w_0, the middle represents w_1, and the lowest dot is w_2. I have not drawn the reflexive arrows, but there should be both a and b-arrows going from each world to itself.

Consider $w_0 +^* p$. The state $w_0(a)$ contains two worlds, w_0 itself, in which p is true, and w_1, in which p is false, so $w_0(a) + p = \{w_0\}$. Applying the definition of $+^*$, this gives us that $(w_0 +^* p)(a) = \{w_0 +^* p\}$.

The state $w_0(b)$ contains only w_0 itself, so $(w_0 +^* p)(b) = \{w_0 +^* p\}$.

This means that $w_0 +^* p$ is that possibility in which both p and q are true, and each agent is fully informed about the world. We could draw the possibility by a single dot with reflexive a and b-arrows. The common ground in $w_0 +^* p$ consists of just a single world: $C(w_0 +^* p) = \{w_0 +^* p\}$.

Consider now the common ground in w_0: $C(w_0) = \{w_0, w_1, w_2\}$. That means that $C(w_0) +^* p = \{w_0 +^* p, w_2 +^* p\}$. Since in $w_0 +^* p$, q is true, and in $w_2 +^* p$, q is false, $w_2 +^* p$ is different from $w_0 +^* p$, and hence, $C(w_0) +^* p$ is different from $C(w_0 +^* p)$. □

I could have chosen a counterexample that is less complex, but w_0 is the most simple example I could find that is reflexive and introspective, and in which the update with p makes sense as the effect of an utterance in a dialogue between b and a. In the following chapter, we will study a simple dialogue game, and the possibility w can be seen to be one where all preconditions for an utterance by b with the sentence p are fulfilled. I refer to section 7.3 for further details.

If one inspects w_0 in the proof of fact 6.10, one sees that a believes that b considers a $\neg q$-world possible only if $\neg p$ is the case. Agent a believes that $p \to \neg\Box_b q$. So, in a sense, the fact that $\neg q$ is a possibility in the common ground in w_0 depends on the fact that a considers $\neg p$ possible, and both a and b know this (although it is not common ground that they do). This is the reason that when one considers the mutual expansion with p, $\neg q$ disappears from the mutual beliefs. This effect is not reflected in the p-expansion of the common ground, because it is not common knowledge that $p \to \neg\Box_b q$.

We get the same kind of result when we choose to represent the common ground in a less detailed way – containing only information about atomic sentences, for example.

Definition 6.11

σ contains *less world-information* than τ, $\sigma \preccurlyeq \tau$, iff for all $v \in \tau$ there is a $w \in \sigma$ such that $w[\mathcal{A}]v$

σ and τ are *atomically equivalent*, $\sigma \approx \tau$ iff $\sigma \preccurlyeq \tau$ and $\tau \preccurlyeq \sigma$ □

Commutativity modulo atomic equivalence (which is in essence the same thing as representing the common ground as a state containing only information about atomic sentences), fails in the same way as it did before:

Fact 6.12 $C(w +^* \phi) \not\approx C(w) +^* \phi$

proof: Use the same counterexample as before. □

As I remarked above, mutual belief may be too strong a notion to use as a model for the common ground. So one may view the property that the expansion of the common ground will never give you any results that are not also mutually believed in the mutually expanded possibility as a minimal correctness condition. In that respect $+^*$ behaves correctly:

Fact 6.13 $C(w +^* \phi) \subseteq C(w) +^* \phi$. □

It turns out that the class of possibilities for which the diagram commutes, modulo $\approx$, coincides exactly with the class of possibilities where $E(w) = C(w)$: the class of possibilities in which 'everybody believes ϕ' implies 'it is mutually believed that ϕ:'

Fact 6.14 $E(w) \approx C(w)$ iff $C(w +^* \phi) \approx C(w) +^* \phi$ for all ϕ □

Because the language we are considering is not very rich in expressive power, we cannot prove a result corresponding to the fact above with $\approx$ interchanged with real identity. We do have the following result:

Fact 6.15 If $C(w) = E(w)$, then $C(w +^* \phi) = C(w) +^* \phi$ □

One can see the fact that the diagram commutes only for possibilities in which everybody's belief is mutual belief as a kind of diagnosis of the problem. In such models, there is no distinction between first-order information (beliefs about the world) and higher-order information (beliefs about beliefs). The only possibilities in which considering the common ground instead of the possibility itself does not lead to loss of information are those in which there is not distinction between first-order and higher-order information.

Other Operators

The trick above, lifting an operator like expansion to a different operator corresponding to a mutual application can easily be generalized: we simply copy the definition for mutual expansion and apply it to an arbitrary function on information states.

Definition 6.16 Let F be any operation on information states. F^* is the following function over possibilities:

$$F^*(w) = v \text{ iff } v[\mathcal{B}]w \text{ and } v(a) = \{F^*(u) \mid u \in F(w(a))\}$$ □

Lifting an operation F to F^* has the following effect: each of the agents applies the operation F to his or her own information state, and then updates the possibilities in the resulting state with F^*. For the formal details I refer to the end of chapter 1.

I will show that assuming that $F^*(C(w)) \preccurlyeq C(F^*(w))$ for each w is inconsistent with assuming that F satisfies the postulates for either contraction or revision, together with the assumption that F is flat:

Definition 6.17 F is *flat* iff for all s and t: if $s \approx t$, then $F(s) \approx F(t)$ □

An operator is flat when an application of F to any two states that are atomically equivalent results in states that are atomically equivalent as well. I think that in general this assumption is not warranted, but when one assumes that F is used to describe change in world-information only (i.e. about information about the values of the propositional variables) the assumption is reasonable: if F expresses change in world-information only, its effects on the world-information should depend on world-information only.

Definition 6.18 (monotony)

An update operator F is *monotone* over an ordering $\leq$ iff it holds that $\sigma \leq \tau$ implies that $F(\sigma) \leq F(\tau)$.

F is *propositionally monotone* iff it is monotone over $\preccurlyeq$ (where $\preccurlyeq$ is the ordering of definition 6.11) □

It turns out when F is flat, monotony of F over $\preccurlyeq$ is a necessary condition for the property that $F^*(C(w)) \preccurlyeq C(F^*(w))$

Fact 6.19 If F is flat, and for each w, $F^*(C(w)) \preccurlyeq C(F^*(w))$, then F is monotone over $\preccurlyeq$.

proof: Assume F is flat, and $F^*(C(w)) \preccurlyeq C(F^*(w))$ for each w. Take any σ and τ such that $\tau \preccurlyeq \sigma$. Take any $v \in \tau$ (assuming that τ is not empty: in that case, $\tau = \sigma$, and we are finished), and define an $\mathsf{S5}$-possibility w as follows:

$$\begin{array}{rcl} w(p) & = & v(p) \text{ for each } p \in \mathcal{P} \\ w(a) & = & \{v \mid \exists u \in \sigma : u \approx v \text{ and } v(c) = w(a) \text{ for all } c \in \mathcal{B}\} \cup \{w\} \\ w(b) & = & \{v \mid \exists u \in \tau : u \approx v \text{ and } v(c) = w(b) \text{ for all } c \in \mathcal{B}\} \cup \{w\} \end{array}$$

Since $w(a)$, and each information state occurring anywhere in $w(a)$, is atomically equivalent to σ, and $w(b)$ and each information state occurring in $w(b)$ is atomically equivalent to τ, it follows that $C(w) \approx \sigma \cup \tau \approx \tau$. Since F is flat, it follows that $F(\tau) = F(C(w))$, and hence that $F^*(\tau) \approx F^*(C(w))$.

By assumption, we know that $F^*(C(w)) \preccurlyeq C(F^*(w))$.

By definition of C, $F^*(w)(a) \subseteq C(F^*(w))$, so $C(F^*(w)) \preccurlyeq F^*(w)(a)$. By definition of F^*, $F^*(w)(a) = F^*(w(a))$ and F^*, $F^*(w(a)) \approx F(w(a))$. Since we have defined w in such a way that $w(a) \approx \sigma$, we have by flatness that $F(w(a)) \approx F(\sigma)$.

Putting all of this together, we get that $F(\tau) \preccurlyeq F(\sigma)$. Since we chose σ and τ arbitrarily, we may conclude that F is monotone over $\preccurlyeq$. □

This result is interesting, because the notions of revision and contraction are not propositionally monotone. Consider for example the following two postulates that have been proposed as conditions on any contraction function:

K–3 If $\sigma \not\models \phi$, then $\sigma - \phi = \sigma$ (vacuity)

K–4 If $\not\models \phi$, then $\sigma - \phi \not\models \phi$ (success)

Fact 6.20 If p is atomic, then any function $-p$ that satisfies (K–3) and (K–4) is not propositionally monotone. □

The proofs of this fact and the following are similar to those given in the next section, section 6.4.

If a revision function satisfies the following two plausible principles, then it is not monotone over $\preccurlyeq$ either.

K*3 If $\sigma \not\models \neg\phi$, then $\sigma * \phi = \sigma + \phi$

K*4 If $\not\models \neg\phi$, then $\sigma * \phi \neq \emptyset$

Fact 6.21 If p is atomic, and $*p$ satisfies to (K*3) and (K*4) then $*p$ is not monotone. □

To recapitulate, we have established that propositional monotony is a necessary condition for the compatibility of internal and external views on information change of common ground. The fact that revision and contraction are not propositionally monotone implies that with respect to these operators, the internal and external view on changes in the common ground do not give the same results.

6.4 A General Framework

I have shown how to formalize our informal picture in possible worlds semantics. In this section I will generalize the results of the previous section by assuming little as possible about the structure of information states, the common ground, or the relation between simple and mutual updates.

We start with some minimal assumptions that we need for representing agents with certain information. First of all, assume there is a set of agents $\mathcal{A}$ and a set of (information-)states $\mathcal{S}$ that those agents may be in. We will assume $\mathcal{A}$ to be finite, lets say $\mathcal{A} = \{1, \ldots, n\}$, and we will use s_i as variables over states that agent i is 'in'. (States may be represented by sets of possible worlds, by sets of sentences, discourse representation structures, databases, situation-theoretical objects, anything that suits your fancy.) We also assume that there is a transitive and reflexive relation $\preccurlyeq$ on $\mathcal{S} \times \mathcal{S}$, similar to the one we defined in the previous section. The idea is that this relation expresses 'containing less information about the world,' i.e. it is a measure of information that disregards information about the epistemic states of the agents. We will write that $s \approx s'$ iff $s \preccurlyeq s'$ and $s' \preccurlyeq s$.

We want to be able to talk about agents having certain information in common, so we need a notion of agents being in certain states together. The simplest way to do this is by representing such a situation by a sequence $\bar{s} = \langle s_0 \dots s_n \rangle$.

Another thing that we assume is that there is some function that extracts the common ground in a situation $\bar{s}$ and that the common ground can be represented by the same kind of object that represents the states of the agents. In other words, we assume that there exists a function C on situations $\bar{s}$ such that $C(\bar{s})$ is a state from $\mathcal{S}$. The following assumption can be seen as a minimal assumption on the function C:

common ground $C(\langle s_0 \dots s_n \rangle) \preccurlyeq s_i$ for any $i \leq n$

We assume that the common ground in a situation contains less information than each of the agents has in that situation. I don't think this is a controversial assumption in any way.

Now take an operation $F : \mathcal{S} \mapsto \mathcal{S}$ and a corresponding notion of a mutual application of this function $F^* : \bar{\mathcal{S}} \mapsto \bar{\mathcal{S}}$ that operates on sequences of states. I will propose a number of assumptions on these functions (all of which were assumed in the previous sections) which together are strong enough to give results similar to those we got in the previous section.

distributivity If $F^*(\langle s_0 \dots s_n \rangle) = \langle t_0 \dots t_n \rangle$, then $F(s_i) \approx t_i$ for all $i \leq n$

To accept this postulate, keep in mind that $\preccurlyeq$ orders states with respect to world-information only. What the assumption says is that if the agents in $\mathcal{A}$ mutually perform the operation F, then their higher-order information may change in all kinds of ways, but the changes in the information they have about the world will be the same as when each of the agents would have applied the operation 'on her own.'

We need a third assumption to guarantee that we have enough states to work with:

fullness We assume that for every two states s and t such that $s \preccurlyeq t$, there is a situation $\bar{s}$ that contains a state t such that $t \approx t'$, and which is such that $C(\bar{s}) \approx s$

This is not a very strong assumption, I believe, but it may help to unravel the definition a little. Fullness says that for any two states s and t such that s contains less world-information than t, there is a situation $\bar{s}$ such that the world-information that is mutually known in $\bar{s}$ is the same as that contained in s, while one of the agents in $\bar{s}$ has the same world-information that is contained in t.[2]

[2] Also for this assumption it is important to note that $\preccurlyeq$ pertains to world information only. Assuming this, I can see no reason for this assumption to fail in any of the representational frameworks that I know of. The proof of fact 6.19 contains a construction of such a state in a possible worlds model.

The last assumption we make is the same as we did before:

flatness If $s \approx t$, then $F(s) \approx F(t)$

Given these four assumptions, we can prove that if our diagram commutes, then F must be monotone over $\preccurlyeq$. In fact, we prove something slightly stronger, corresponding to fact 6.19, namely that monotony is a necessary condition for $F(C(\bar{s})) \preccurlyeq C(F^*(\bar{s}))$:

Fact 6.22 Assume that the four properties formulated above hold. Then it also holds that if $F(C(\bar{s})) \preccurlyeq C(F^*(\bar{s}))$, then F is monotone over $\preccurlyeq$.

proof: Take any s and t such that $t \preccurlyeq s$. Since $\mathcal{S}$ is full, we can find $\bar{s} = \langle s_0 \dots s_n \rangle$ such that $s \approx s_i$ for some $i \leq n$ and $C(\bar{s}) \approx t$. Since F is flat, $F(t) \approx F(C(\bar{s}))$. By assumption, $F(C(\bar{s})) \preccurlyeq C(F^*(\bar{s}))$.

Let $F^*(\bar{s}) = \langle t_0 \dots t_n \rangle$. We assumed that $C(F^*(\bar{s})) \preccurlyeq t_i$. By distributivity, $t_i \approx F(s_i)$, and using flatness again, we have that $F(s_i) \approx F(s)$.

Since we assumed that $\preccurlyeq$ is transitive, we can combine these observations and conclude that $F(t) \preccurlyeq F(s)$.[3] □

Since none of the operations considered above was originally defined to be applied to such abstract objects as the states introduced above, we still need to show that this abstract result applies to contraction or revision functions. Of course, we will not be able to prove anything about the original notions of expansion, contraction and revision. Instead, I will reformulate some postulates yet again (in fact I will slightly weaken some), and then prove how failure of monotony follows from them.

To show that contraction functions are not monotone over $\preccurlyeq$, we need to reformulate the postulates for contraction in such a way that they apply to states in general. And for doing that, we need to extend our ontology: we need a language and a relation of $\models$ of 'acceptance' between $\mathcal{S}$ and the language. Think of $s \models \phi$ as meaning that ϕ is accepted in state s, that the information that ϕ is subsumed by the information in s, or that the information that ϕ is contained in s.

Consider the following postulates for contraction:

K–2′ $s - \phi \preccurlyeq s$

K–3′ If $s \not\models \phi$, then $s - \phi = s$ (vacuity)

K–4′ If ϕ is not a tautology, then $\sigma - \phi \not\models \phi$ (success)

The original formulation of (K–2′) from Alchourrón et al. (1985) uses a stronger ordering instead of $\preccurlyeq$, so (K–2′) is a weaker version of the postulate. (K–3′)

[3] I have skipped over matters pertaining to the possible partiality of the function F. If one defines monotonicity as a property that need only hold for values on which F is defined, the proof will work just as well.

is exactly the same as the original definition, which I also used in the previous section. (K–4′) introduces the notion of a 'tautology:' this may be taken as a primitive notion, or it may be taken as defined as 'being accepted in each state' or as 'being accepted in the minimal state.'

To show a function $-\phi$ satisfying these three postulates is not monotone over ϕ, if ϕ is not a tautology, we need to be sure that $(\mathcal{S}, \preccurlyeq)$ contains enough structure.

We will assume that there are states s and t in $\mathcal{S}$ such that $s \not\models p$, $t \not\models p$, and for all u such that $s \preccurlyeq u$ and $t \preccurlyeq u$, $u \models p$. Moreover, we assume that there is in fact a u such that $s \preccurlyeq u$ and $t \preccurlyeq u$. (For an intuitively acceptable example, consider states s and t such that $s \models q$, $t \models q \rightarrow p$.)

If $(\mathcal{S}, \preccurlyeq)$ has these properties, we can prove that $-p$ is not monotone over $\preccurlyeq$. For assume that $-p$ is monotone. We know that $s - p = s$ and $t - p = t$, by (K–3). Now take any u that contains more information than both s and t. By monotony, $u - p$ must contain more information than both s and t. But by assumption, every such state is one in which p is accepted, contradicting (K–4′).

The postulates for revision presuppose that we have negation in our language, and that $\mathcal{S}$ contains an inconsistent state $\perp$. We will assume that if s is a state such that $s \models p$ and $s \models \neg p$ for some sentence p, then $s \approx \perp$. Consider the following postulates:

K*2 $s * \phi \models \phi$.

K*3′ If $s \not\models \neg\phi$, then $s \preccurlyeq s * \phi$.

K*4′ If $\neg\phi$ is not a tautology, then $\sigma * \phi \not\approx \perp$.

The postulate (K*2) is just the original postulate from Alchourrón et al. (1985). It is not hard to see that (K*3′) is a weakening of (K*3), assuming at least that $s \preccurlyeq s + \phi$. Similarly, we have weakened (K*4) to the effect that if ϕ is not a contradiction, then a revision with ϕ will not be atomically equivalent with the inconsistent state.

Let p be such that $\neg p$ is not a tautology, and assume we have states s and t such that $s \not\models \neg p$ and $t \not\models \neg p$, and for all u such that $s, t \preccurlyeq u$, $u \models \neg p$. Assume moreover that there exists such a u. (Consider, e.g., $s \models q \rightarrow \neg p$, $t \models q$, similar as before.) Take any u such that $s, t \preccurlyeq u$. It holds, by (K*3′), that $s \preccurlyeq s * p$, and by monotony, that $s * p \preccurlyeq u * p$. Similarly, $t \preccurlyeq t * p \preccurlyeq u * p$. But then, by assumption, $u * p \models \neg p$. But according to (K*2), $u * p \models p$, from which it follows that $u \approx \perp$, which contradicts (K*4′).

6.5 Conclusions

In this chapter, I have compared two ways of modeling changes in the common ground. We can model changes in the common ground from an external point of view, by defining a function of information change that operates on representations of the common ground directly, or we can model such change internally, by viewing changes in the common ground as derived from changes in the belief states of the participants involved.

The main conclusion to draw from the results of this chapter is that if we identify the common ground with that which is mutually believed, the internal and the external viewpoints are incompatible with each other. This holds even for simple notions of information change such as expansion.

If we consider expansion, then the external way of modeling changes in the common ground may give results that are too weak when compared with the internal view. If this discrepancy is a problem at all, I don't think it is a very serious one. First because it seems that one does not need all mutual beliefs to be in the common ground anyway. But also because one of the reasons of using a separate representation of the common ground is that it is a less complicated way of modeling dialogue than keeping track of the states of the participants; this means loosing certain information about the relations between world-information and higher-order information, but fact 6.13 shows that this is basically harmless when considering expansion.

The result that a function that is flat has to be monotone for the changes in the common ground to be mutual beliefs, and that neither contraction nor revision are monotone seems to be a more serious problem. On the other hand, both notions are notorious for their indeterminacy. What the results seem to say is that if one uses a simple deterministic function to model contraction or revision of the common ground, it may be that the resulting common ground will contain information that is not mutually believed. But if one takes a more lenient view on contraction or revision, and sees the operations as processes that involve some more or less arbitrary decisions on what kind of information to discard, i.e. if one considers the result of a revision process as, to a certain extent, unpredictable, it will be unclear in general what exactly is in the resulting common ground, and it will be even less clear to each of the participants what is mutually believed (since the latter involves reasoning about the belief change of the other agents, and their reasoning about each other's belief change, etcetera). The negative results seem to give just another argument that revision and contraction processes are not to be modeled by deterministic functions.

7

Dirt, Dimes and Dialogue

This chapter consists of three parts. In the first two parts, I will show how DEL can be used to analyze two classical puzzles that are about belief and belief change. The first puzzle I will discuss is known under the name of the 'puzzle of the dirty children.' The analysis that I will give was already discussed in Gerbrandy (1994) and in Gerbrandy and Groeneveld (1997). The second puzzle is known as the 'surprise exam paradox.' These two puzzles can be represented in DEL in a natural way. As we will see, it turns out that some of the more puzzling aspects of these puzzles can be explained as being connected to the fact that unsuccessful updates play an important role in them.

In the third section of this chapter, I will define a simple dialogue game. It is meant as a first sketch of how DEL can be used as a tool to study certain aspects of dialogue.

7.1 The Dirty Children

The puzzle of the dirty children is an example of a puzzle that involves sophisticated reasoning about knowledge. The description of the puzzle of the dirty children that I give here is adapted from Barwise (1981):

There is a group of children playing together. During their play some of the children, say k of them, get mud on their foreheads. Each can see the mud on others but not on his own forehead. Along comes a father, who says, "At least one of you has mud on his head." He then asks the following question, over and over: "Can any of you prove that you have mud on your head?" Assuming that all the children are perceptive, intelligent, truthful, and that they answer simultaneously, what will happen?

There is a "proof" that the first $k-1$ times the father asks the question, the children will all say "no" but that the k-th time the children that are dirty will answer "yes."

The proof is by induction on the number of dirty children k. For $k = 1$ the result is obvious: the dirty child sees that no one else is muddy, so he must be the muddy one. If there are two dirty children, say a and b, each answers "no" the first time, because of the mud on the other. But, when b says "no," a realizes he must be muddy, for otherwise b would have known the mud was on his head and answered "yes" the first time. Thus a answers "yes" the second time. b goes through the same reasoning. Now suppose there are three dirty children, a, b, c. Child a argues as follows. Assume I don't have mud on my head. Then, by the $k = 2$ case, both b and c will answer "yes" the second time. When they don't, he realizes that the assumption was false, that he *is* muddy, and so will answer "yes" on the third question. Similarly for b and c.

This puzzle occurs under different guises in the literature: it is a variant of the puzzle of the cheating husbands (see for example Moses et al., 1986), the wise men puzzle (in e.g. McCarthy, 1990) and the Conway paradox (e.g. Emde Boas et al. (1980)). Parikh (1992) describes an intriguing infinitary version of the Conway paradox.

There are several aspects of the puzzle that need explanation. First, there is the role that the father's first utterance plays in the puzzle. When $k > 1$, it is already known by each of the children that at least one of them is dirty, since each of them can see another child that is dirty. In this case, the father tells the children something they already know. Yet, the proof breaks down if we omit the father's first announcement from the description of the puzzle. That means that the children must learn something from the father's utterance that they did not know before. A proper analysis of the puzzle should explain what this 'something' is.

The role of the negative answers of the children is similar. Consider the case where three children are dirty. Since each child a can see at least two other dirty children, she can infer that, no matter whether a herself is dirty, each of the children can see at least one child that is dirty. From this, a can safely conclude that none of the children know whether they are dirty. But that means that a already knows that all children will answer 'no' after the first time the father asks his question.

So the father's first announcement, and many of the children's answers, seem to give information that is already known by all of the children. At the same time, for the 'proof' to work it is necessary to assume that these utterances were made. An analysis of the puzzle should explain what exactly the contribution of these utterances is.

Another surprising aspect of the answers of the children is that when $k > 2$, the children give the same answer to the same question over and over again. Yet, any time they answer 'no,' something changes in the situation described; otherwise there would be no point in repeating the same answer over and over. So, another aspect of the puzzle that needs clarification is the question how repeating the same 'sentence' can have considerable effect each time it is uttered.

We can slightly alter the description of the puzzle to sharpen this point. Suppose the father announces, over and over again, "None of you knows that you are dirty," instead of asking a question and getting an answer. It is not hard to see that the proof works just as well in this case. But now, we seem to have a case where uttering the same sentence over and over again can have a considerable effect each time. Even more surprising, perhaps, is that if there are k dirty children, some of the children *do* know that they are dirty after the father has announced $k-1$ times that none of the children knows that he is dirty.

In the rest of this section, I will first show how the puzzle can be formalized in DEL, and follow up with a discussion of how the aspects mentioned above are modeled, and partly explained, by this analysis.

Formalization

I will show how one can formalize the description of the puzzle, and the reasoning involved, in dynamic epistemic semantics. I will give a formal proof of the fact that if k children that are dirty, then after $k-1$ answers to the father's question, the children that are dirty know that they are dirty.

Let $\mathcal{A}$ be a set of children playing in the mud. Consider a language that contains a propositional atom p_a for each $a \in \mathcal{A}$. If p_a is true, that means that a is dirty. I start by introducing some convenient abbreviations:

- Each child can see the forehead of each of the other children. So, if a child is dirty, each of the other children knows that she is dirty. This can be expressed by the conjunction of all sentences of the form $(p_a \to \Box_b p_a) \wedge (\neg p_a \to \Box_b \neg p_a))$ for each a and b in $\mathcal{A}$ such that $a \neq b$. This conjunction is abbreviated as by vision.

- It is common knowledge among all children that each forehead can be seen by each of the other children, i.e. the sentence vision is common knowledge among all children. We can express this as $C_{\mathcal{A}}\mathsf{vision}$.

- Before asking the children whether they know if they are dirty or not, the father announces in front of all children that at least one of them has a dirty forehead. Let father be the sentence $\bigvee\{p_a \mid a \in \mathcal{A}\}$, which expresses that at least one of the children is dirty.

- After the father's announcement, all children answer the question 'Do you know whether you are dirty or not?' The children answer either 'yes' or 'no.' Let no be the sentence $\bigwedge\{(\neg\Box_a p_a \wedge \neg\Box_a \neg p_a) \mid a \in \mathcal{A}\}$, which is the sentence that expresses that none of the children knows that she is dirty (i.e. the information expressed by all children answering 'no' at the same time).

- Finally, let for each $\mathcal{B} \subseteq \mathcal{A}$, $\mathsf{dirty}(\mathcal{B})$ abbreviate the formula $\bigwedge_{b \in \mathcal{B}} p_b \wedge \bigwedge_{b \notin \mathcal{B}} \neg p_b$. This formula expresses that all and only the children in $\mathcal{B}$ have dirty foreheads.

We now have to make precise what exactly the effect is of the father saying the sentence father, and what the effect is of the children saying no.

The situation in the puzzle is highly idealised. I will assume that the effect of a public announcement in such an idealised situation is twofold: each of the children updates with the sentence uttered, and each of the children knows that the sentence has been publicly announced. In short, we will model the change brought about by the father's utterance of father as the meaning of the 'program' $U_{\mathcal{A}}?\mathsf{father}$, and the effect of the children answering 'no' as $U_{\mathcal{A}}?\mathsf{no}$. This assumption is so important that I highlight it in the following definition. We will discuss this definition in more detail in section 7.3.

Definition 7.1 (public utterance)
The effect of a public utterance of a sentence ϕ in a group $\mathcal{B}$ can be expressed by the program $U_{\mathcal{B}}?\phi$. □

We have now all the ingredients we need for a formal description of the puzzle. We can now formally show the following. Suppose w is a possibility where exactly k children are dirty, and where it is commonly known among all children that they can see each other. If w is updated with a public announcement (by the father) that at least one of the children is dirty, followed by $k-1$ public announcements that none of the children know that they are dirty (the representation of the father's question and all children answering 'no'), the resulting state is one where all dirty children know that they are dirty. Formally expressed, this boils down to the following statement:

Proposition 7.2 Let $\mathcal{B}$ be a set containing exactly k children ($k \geq 1$). We let $[\pi]^m$ stand for the sequence $[\pi] \ldots [\pi]$ of m updates with $[\pi]$.

It holds for all $a \in \mathcal{B}$ that:

$$\mathsf{dirty}(\mathcal{B}), \mathsf{vision}, C_{\mathcal{A}}\mathsf{vision} \models [U_{\mathcal{A}}?\mathsf{father}][U_{\mathcal{A}}?\mathsf{no}]^{k-1}\Box_a p_a$$

proof: I will give a syntactical proof of this statement, using the axioms of DEL that were formulated in chapter 4. Although we do not have a complete axiomatization for the language containing the common knowledge operators $C_{\mathcal{B}}$, it is easy to see that the following axiom is sound, if $b \in \mathcal{B}$:

$$\vdash C_{\mathcal{B}}\phi \to \Box_b(\phi \wedge C_{\mathcal{B}}\phi)$$

I will use of this axiom in the proof.

In the proof, will write [father] instead of [$U_{\mathcal{A}}?\mathsf{father}$], and [$\mathsf{no}$] for the operator [$U_{\mathcal{A}}?\mathsf{no}$].

The proof of the proposition is by induction on the number k of dirty children. Assume first that only one child, say a, is dirty. In classical modal logic, it holds that if a is the only dirty child, and a can see all other children, then a knows that if at least one child is dirty, it must be herself:

$$\mathsf{dirty}(\{a\}), \mathsf{vision} \vdash \Box_a(\mathsf{father} \rightarrow p_a)$$

We use axiom 5 to conclude that:

$$\mathsf{dirty}(\{a\}), \mathsf{vision}, C_{\mathcal{A}}\mathsf{vision} \vdash \Box_a(\mathsf{father} \rightarrow [\mathsf{father}]p_a)$$

whence, by axiom 6

$$\mathsf{dirty}(\{a\}), \mathsf{vision}, C_{\mathcal{A}}\mathsf{vision} \vdash [\mathsf{father}]\Box_a p_a$$

For the induction step, let $\mathcal{B}$ be a set of $k+1$ children, and $a, b \in \mathcal{B}$. It holds by induction hypothesis and the necessitation rule that if $\mathcal{B}'$ has k elements and $b \in \mathcal{B}'$, then

$$\vdash \Box_a((\mathsf{dirty}(\mathcal{B}') \wedge \mathsf{vision} \wedge C_{\mathcal{A}}\mathsf{vision}) \rightarrow [\mathsf{father}][\mathsf{no}]^{k-1}\Box_b p_b)$$

Since

$$C_{\mathcal{A}}\mathsf{vision} \vdash \Box_a(\mathsf{vision} \wedge C_{\mathcal{A}}\mathsf{vision})$$

and

$$\mathsf{dirty}(\mathcal{B}), \mathsf{vision} \vdash \Box_a(\neg p_a \rightarrow \mathsf{dirty}(\mathcal{B}/\{a\}))$$

it follows for $a, b \in \mathcal{B}$ such that $a \neq b$ that:

$$\mathsf{dirty}(\mathcal{B}), \mathsf{vision}, C_{\mathcal{A}}\mathsf{vision} \vdash \Box_a((\neg p_a \wedge \mathsf{father}) \rightarrow [\mathsf{father}][\mathsf{no}]^{k-1}\Box_b p_b)$$

from which it follows, using axioms 3 and 5, and the fact that $\vdash \mathsf{no} \rightarrow \neg\Box_b p_b$, that

$$\mathsf{dirty}(\mathcal{B}), \mathsf{vision}, C_{\mathcal{A}}\mathsf{vision} \vdash \Box_a(\mathsf{father} \rightarrow [\mathsf{father}][\mathsf{no}]^{k-1}(\mathsf{no} \rightarrow p_a))$$

By axiom 6, then

$$\mathsf{dirty}(\mathcal{B}), \mathsf{vision}, C_{\mathcal{A}}\mathsf{vision} \vdash [\mathsf{father}]\Box_a[\mathsf{no}]^{k-1}(\mathsf{no} \rightarrow p_a)$$

and using the lemma below, finally,

$$\mathsf{dirty}(\mathcal{B}), \mathsf{vision}, C_{\mathcal{A}}\mathsf{vision} \vdash [\mathsf{father}][\mathsf{no}]^k\Box_a p_a$$

Lemma 7.3 For each m: $\Box_a[\mathsf{no}]^k(\mathsf{no} \rightarrow p_a) \vdash [\mathsf{no}]^{k+1}\Box_a p_a$

This is proven by induction on k. If $k = 0$, then $\Box_a(\mathsf{no} \to p_a)$ is equivalent by axiom 5 to $\Box_a(\mathsf{no} \to [\mathsf{no}]p_a)$, which is, by axiom 6, equivalent to $[\mathsf{no}]\Box_a p_a$.

For the induction step, assume that:

$$\Box_a[\mathsf{no}]^{k+1}(\mathsf{no} \to p_a)$$

This implies that

$$\Box_a(\mathsf{no} \to [\mathsf{no}]^{k+1}(\mathsf{no} \to p_a))$$

From which it follows by axiom 6 that

$$[\mathsf{no}]\Box_a[\mathsf{no}]^k(\mathsf{no} \to p_a)$$

whence, by induction hypothesis, that

$$[\mathsf{no}][\mathsf{no}]^{k+1}\Box_a p_a$$

This completes the proof of proposition 7.2.[1] $\Box$

Discussion

The puzzle of the dirty children and related puzzles have been discussed relatively extensively in the literature, and several formalizations have been given. I believe that the analysis presented here adds to earlier approaches in an essential way.

First of all, the informal description of the puzzle has been rephrased in the object language of an independently motivated logic with fairly little tinkering: the utterances of the father and the children translate straightforwardly into the object language; the only assumption that is not trivial is that the effect of a public announcement is modeled by the $U_{\mathcal{A}}$-operator. Such an analysis has not been given before: all earlier formalizations of the puzzle that I know of consist of a more or less *ad hoc* model of the information and information change involved in the puzzle. That means that each variant of the puzzle calls for a new analysis and the construction of a new model. The relatively straightforward way in which the puzzle can be formalized in DEL suggests that similar problems may be formulated in the same way. We can easily model variants of the scenario in a straightforward way. For example, we could let the father announce that at least three children are dirty instead of at least one. In this scenario, the dirty children would know that they are dirty after $k-3$ answers of 'no.' A slightly greater deviation of our scenario is the 'three wise men puzzle' of McCarthy (1990), which is similar to the dirty children puzzle, except that the children answer the question whether they are dirty one by one, instead of all at the same time. We can straightforwardly

[1] I have not proven that the children do not know that they are dirty *before* they have answered the question $m-1$ times. To show that, one needs an extra assumption: in the initial possibility, none of the children knows whether she is dirty or not, and this fact is common knowledge. A proof can then be given along the lines of the proof given here.

model this version in DEL. In the section after this, I will show how one can analyze the surprise exam paradox in DEL.

The fact that our formalization of the puzzle gives results similar to Barwise's semi-formal results shows that the paradoxical flavor of the puzzle does not stem from a logical mistake. This strongly suggests that the discrepancy between the ideal situation described in the puzzle and a 'real life' situation should not be explained as a difference in principles of logic, but as a result of the complexity of the reasoning involved in the puzzle and the way it depends on the strong trust the children have in each other's reasoning capabilities.

Consider now the problems that were identified in the introduction of this section. First, we needed to clarify the role of the utterances of sentences whose truth is already known, such as the father's statement, which all children know to be true if there are two or more dirty children present. In such a situation, each of the children already knows that one of the children is dirty (since everyone can see a dirty child). The formal correlate of this fact is a theorem; it holds for each a, b and c such that $a \neq b$ that:

$$p_a \wedge p_b, \mathbf{vision} \vdash \Box_c \mathsf{father}$$

The point of the father's statement is that his announcement has the effect of a mutual update. The effect of a mutual update with father is that it becomes common knowledge that at least one child is dirty, which was not the case before: $p_a \wedge p_b, \mathsf{vision}, C_{\mathcal{A}}\mathsf{vision} \not\vdash C_{\mathcal{A}}\mathsf{father}$. If there are two children that are dirty, it is indeed the case that each of them knows that at least one of them is dirty (they can see the other dirty child). But they do not know of each other that they know this. For example, child a, not knowing whether she herself is dirty, cannot be sure that b can see a dirty child (if a is clean, b sees only clean foreheads).

A similar fact holds for the sentence no. If $k \geq 3$, then, immediately after the father's first announcement, each of the children knows that no is true. But again, it is not the case that no is common knowledge. For example, if $k = 3$, and a is a dirty child, she can see two other dirty foreheads. Say one of these foreheads belongs to b. Since a does not know whether she herself is dirty, she believes it may be possible that b sees just a single dirty forehead, that of c. Since neither b knows of himself whether he is dirty or not, he cannot be sure that c knows that c is dirty: if b is clean, c would (in b's reasoning under a's assumption that she is clean herself) know that he is dirty. So, a believes it is possible that b believes it is possible that c knows that c is dirty. Which means that it is not common knowledge that c does not know he is dirty, and therefore, neither is the sentence no common knowledge.[2]

[2] And not, as Landman (1986) writes about the Conway paradox, which is a variation on the dirty children puzzle: "The point is: they know from the start that the first answer will be "no." They also know that the second answer will be "no," and the third. Hence, they could update their information *before* the first round was played three, respectively four times." The point

These observations are not new, but our analysis adds to earlier ones in that it is now possible to formulate such facts in the object language.

One remaining aspect of the puzzle highlighted in the present analysis is the fact that the children keep on saying ‘no’ until ‘suddenly’ some children answer ‘yes.’ This fact suggests that each answer supplies new information, although, ‘syntactically,’ the children say the same thing each time. This is directly reflected in our semantics: an update with no changes the possibility in a certain fixed way, resulting in a new possibility in which another update with no may change the possibility again.

This is an example of the failure in DEL of the following principle of *success*, that we discussed in chapter 4:

$$\text{if } b \in \mathcal{B} \text{ then } \models [U_{\mathcal{B}}\phi]\Box_b\phi$$

This states that after a group update with a sentence ϕ, each member in the group knows that ϕ. This is *not* a property of updates in general, and the example of no suggests that this is right.

For this particular example, it even holds that there are possibilities w such that $w \models [U_{\mathcal{A}}?\mathsf{no}]\Box_b\neg\mathsf{no}$. This is not as surprising as it may seem. At the moment where the children have answered ‘no’ $k-2$ times, all the dirty children know that if any of the other children do not know they are dirty, then they must be dirty themselves. As soon as they learn that the other children do not know they are dirty, they conclude that they themselves must be dirty.

This observation has interesting repercussions for a theory of dialogue. In an idealized situation such as described in the puzzle, one would expect that the effect of an utterance of an indicative sentence on the common ground between the participants is very simple: the information expressed by the sentence uttered is simply added to the common ground. This is corroborated by the effect of the father’s first announcement: the information contained in that utterance becomes common knowledge after the utterance. More precisely, if one of the dialogue participants utters a sentence ϕ, and none of the other participants has reason to disagree, ϕ will be common knowledge after the utterance. In fact, this is what is proposed by, among others, Stalnaker (1978).

However, for a right understanding of the puzzle, the effect of an utterance cannot be as simple as this. This is shown by the observation that after all children say “no,” it is *not* common knowledge that none of the children know whether they are dirty. Quite the contrary, the puzzle shows that the result may even be a situation in which some of the children (the dirty ones) *do* know that they are dirty.

---

is that the information they get from the answer “no” is not that none of the children knows that he is dirty: that they already know. The crucial information they get is that the children *mutually learn* that the answer is negative. Paraphrasing Landman, we can say that they can update their information with the fact that the answer *will be* “no” before the exchange took place, but they cannot update with the information that the answer “no” *has been given*.

In our analysis, the effect of an utterance is described as a 'mutual update.' Roughly, this means that the effect of an utterance of a sentence ϕ (in an idealized case, such as in the puzzle, where all dialogue participants are 'perceptive, truthful, intelligent') is that all dialogue participants learn that ϕ, all participants learn that all participants learn that ϕ, they all learn that all of them have learned that ϕ, and so on. We have already seen in chapter 4 that in most cases a mutual update with a sentence ϕ will result in a situation in which ϕ is common knowledge. Only when a sentence expresses that certain dialogue participants *lack* certain information, it may happen that the sentence is not common knowledge after a mutual update with that sentence; in fact, it may even happen that the negation of that same sentence becomes common knowledge, as the puzzle shows.

To sum up: the notion of a mutual update seems to be the right analysis of the effect of the children answering 'no' in the muddy children problem. At the same time, it fits well with our intuition that, in most cases, the effect of an utterance is that the sentence uttered becomes common knowledge.

7.2 The Surprise Examination Paradox

We now turn to a discussion of another puzzle that is well-known from the literature. It is probably best known under the name of the 'surprise examination paradox.' The puzzle is discussed in several places. The first published formulation is from O'Connor (1948). Another quite concise version is that of Quine (1953). The following version is derived from that of Landman (1986).

A series of n numbered boxes is opened in sequence by the quiz master, starting from number 1, then number 2, etcetera. One of the boxes contains an enormous amount of money, and the quiz master knows which box it is. A player, b, gets the money if he *knows*, just before the box containing the money is opened, that this box is the one with the money in it. Player b is not allowed to guess; he must have a convincing argument that the money is in the box to be opened.

Suppose that the box with the money in it is somewhere in the middle – say there are five boxes, and the money is in the fourth. In that case, player b will never win the game, because at the moment that the fourth box is opened, he has no reason to assume that box 4 is not an empty box. Since the quiz master knows which box contains the money, she knows that b cannot win the game.

The 'paradox,' now, is the following. Suppose the quiz master say to b: "You cannot win the game." As we have seen, this is true. Now b reasons as follows. "Suppose the money is in the last box. In that case, I would know that the money is in that box at the moment when all other boxes were opened, and I would win the game. So, if the quiz master tells the truth, the last box is empty. But if this is true, the money cannot be in box 4 either, because I know (now) that the last box is empty, and so, if boxes 1 to 3 were opened, the money had to be in the fourth box, and I would win the game as well. I can repeat this proof for all boxes. But then I have to conclude that all boxes are empty. This is in contradiction with what I know of the game. Therefore the quiz master must be lying to me."

In contrast to the puzzle of the dirty children, the surprise examination paradox has the flavour of a real paradox. The conclusion in the puzzle of the dirty children is just unlikely (it is not very plausible that children are able of such sophisticated reasoning, and the faith that they are presumed to have in each other's capabilities is an even stronger idealization). In the surprise examination paradox, the conclusion is perhaps not a clear-cut contradiction, but very close to it: the quiz master is telling the truth, b uses the quiz master's announcement together with some other patently true facts as premisses for a simple inductive proof, but the conclusion of this proof is false.

I believe that a great deal of confusion about the surprise examination scenario lies in the fact that the quiz master's statement "You cannot win" is ambiguous between a reading that can be roughly paraphrased as "You cannot win, given (the information you have in) the situation *as it is at the present moment*" and "You cannot win, given the situation right now, and *neither will you win at some moment in the future*, when you get more information."

I will first discuss the second paraphrase of the announcement of the quiz master, and then concentrate on the first paraphrase, which I consider the be the most interesting in the present context.

Suppose the quiz master announces to b "You cannot win the game, now or in the future." Suppose moreover that the quiz master starts the game immediately after his announcement: the only information b has about the game is that exactly one box is full of money, and that the quiz master told him he cannot win, now or in the future. If player b believes the quiz master, he can reason as we described above and conclude that his information is inconsistent. Note that there is no paradox here yet: the same would happen if the quiz master would have announced a contradiction, or simply something that b knew to be false.[3] On the other hand, if b does not believe what the quiz master says, the quiz master's announcement will have no effect on his information about the contents of the boxes, and therefore he will not win the game. But this makes the quiz master's announcement true.

In the second reading, the quiz master's announcement is very similar to more standard self-referential epistemic 'paradoxes' such as

"This very sentence is true iff you do not believe it."

If you believe that this sentence is true, you can infer that it must be false. So, if you believe this sentence is true, your information is inconsistent. On the

[3]There is the further 'meta-complication' that, since b's information is inconsistent, he cannot know for sure which box contains the money. So, the quiz master would be right after all with his prediction that b cannot win. But let us disregard this.

other hand, if you believe that the sentence is false, you do not believe that it is true. But then the sentence must be true after all. Examples such as these are interesting, but they have been discussed elsewhere (cf. Shaw (1958) for an analysis of the surprise examination paradox along these lines) and are not the topic of this dissertation. It should be possible to explicate the second reading in precise formal detail using the framework of Fagin et al. (1995), where epistemic logic is combined with a logic of time.

Paraphrased in this second way, the formulation of the paradox loses much of its force. One of the most salient features of the 'paradox' is that the player derives an inconsistency from statements that are obviously true. In the second reading, the announcement of the quiz master is *not* simply true, and the paradox loses most of its poignancy.

When the money is not in the last box, the first paraphrase of the announcement of the quiz master ("you cannot win given what you know now") is a sentence that *is* true. I will concentrate on giving an analysis of this first paraphrase in the following, because it is with respect to this sentence that an analysis using DEL can throw new light on the paradox of the surprise exam.

Let us start with formalizing the main ingredients of the puzzle.

Let n be the number of boxes. Each of the n boxes can be either full or empty. We can represent this in the object language by choosing, for each $i \leq n$, a propositional variable p_i to represent the proposition that the i-th box contains the money.

To keep matters simple, we will leave the quiz master out of consideration, and model only the information of the player, b. So, the puzzle will be modeled by a possibility with only a single agent, b.

First of all, b knows that exactly one box is filled. Formally, this means that the following sentence, abbreviated as 'onebox', is true, and known to be true by b:

$$\mathsf{onebox} = \bigvee_{1 \leq i \leq n} (p_i \wedge \bigwedge_{j \neq i} \neg p_j)$$

We can represent the statement that b wins the game by the following conjunction:

$$\mathsf{win} = (p_1 \to \Box_b p_1) \wedge (p_2 \to [U_b \neg p_1]\Box_b p_2) \wedge \ldots \wedge (p_n \to [U_b \neg p_1] \ldots [U_b \neg p_{n-1}]\Box_b p_n)$$

Transcribed in a sort of English, this sentence expresses that if box 1 contains the money, then b should know this already to win the game; if box 2 contains the money, b knows this after learning that box 1 is empty (after the first box has been opened); that if box 3 contains the money, b will know this after learning that the previous boxes are empty; etcetera.[4]

The 'paradox' now lies in the fact that in K45, it holds that b's information is inconsistent if he believes both that exactly one box is filled with money, and that he cannot win:

Proposition 7.4 $\models_{\mathsf{K45}} (\Box_b \mathsf{onebox} \wedge \Box_b \neg \mathsf{win}) \rightarrow \Box_b \bot$

proof: A deduction in DEL, following the informal reasoning in the paradox. First note that $\neg\mathsf{win}$ implies, under the assumption that onebox is true, that

$$p_1 \rightarrow \neg\Box_b p_1 \wedge (p_2 \rightarrow \neg[U_b p_1]\Box_b p_2) \wedge \ldots \wedge (p_n \rightarrow \neg[U_b p_1]\ldots[U_b p_{n-1}]\Box_b p_n)$$

Some more reasoning shows that

$$\vdash_{\mathsf{DEL}} \Box_b(\mathsf{onebox} \rightarrow (p_n \rightarrow [U_b\neg p_1]\ldots[U_b\neg p_{n-1}]\Box_b p_n))$$

This means that b knows that if the last box is filled, then he will win the game. Since $\Box_b\neg\mathsf{win}$ and $\Box_b\mathsf{onebox}$, it follows that $\Box_b\neg p_n$: b knows that the money is not in the last box.

From the fact that $\Box_b\neg p_n$, we can derive in DEL that

$$\Box_b(p_{n-1} \rightarrow [U_b\neg p_1]\ldots[U_b\neg p_{n-2}]\Box_b p_{n-1})$$

This formula expresses that b knows he will win the game if the money is in box $n-1$. Using the premiss that b knows that $\neg\mathsf{win}$ another time, it follows that $\Box_b\neg p_{n-1}$. Repeating the argument some more times leads to the conclusion that $\Box_b\neg p_i$ for each $i \leq n$, and together with $\Box_b\mathsf{onebox}$, we conclude that $\Box_b\bot$. $\Box$

There is no real paradox here yet. But we do have a phenomenon that needs explanation. The point is that if the money is not in the last box, both sentences $\neg$win and onebox are true, but if b believes that both of these sentences are true, he can derive an inconsistency. But even the observation that b can derive an inconsistency from true sentences is no reason to call the puzzle a 'paradox.' There are other examples of sentences that can be true but impossible to believe for certain agents. Consider the following version of Moore's paradox: "The cat is on the mat but b does not believe it." This sentence is perfectly consistent, but b cannot consistently believe both that the cat is on the mat and that he does not believe it. A sentence of the form $p \wedge \neg\Box_b p$ is an example of a sentence that can be true, while it cannot be believed by b.

Sentences such as this have always been a bit of a problem in classical logic. Even if saying things like "The cat is on the mat but you don't believe it" is not a very good way to package the information you want to convey, the sentence can convey information that may very well be true, and the hearer should be able

[4]Note that this sentence is equivalent to the sentence $(p_1 \rightarrow \Box_a p_1) \wedge (p_2 \rightarrow \Box_a(\neg p_1 \rightarrow p_2) \wedge \ldots \wedge (p_n \rightarrow \Box_a((\neg p_1 \wedge \ldots \wedge \neg p_{n-1}) \rightarrow p_n))$

to understand this and incorporate the new information into the information she already has. In the classical logic of belief, it is hard to see how this could be done. There seem to be only two options open, both unsatisfactory. The hearer either adds the new sentence to her already existing stock of beliefs, in which case her beliefs become inconsistent (in K45), or she does not add the sentence to her existing beliefs, in which case she does not make very efficient use of the information offered to her. DEL provides us with a middle way, as we have seen in chapter 4 and in the previous section. Updates with sentences that express lack of information of the agent who learns the sentence are not always successful: an agent can learn that such a sentence is true without coming to believe that the sentence is true.

The formula win that I have used to express the statement that player b will win the game can be seen as a complicated version of Moore's paradox. To see this, it is useful to consider the puzzle in more detail.

Suppose there are five boxes, and that the money is in the fourth box. Clearly, whether b will win the game or not depends on what b knows about the location of the money. In the initial situation, as far is b's information goes, the money may be in any one of the boxes. This means that if the money is not in the last box, then b will not win the game. The quiz master makes her announcement on the basis of b's information about the situation, at that moment, together with her own knowledge of the fact that it is box 4 that contains the prize. The effect of her announcement, however, is that b learns something new about the game: the fact that he cannot win implies that the money is not in the last box. But now the situation has changed: before the announcement of the quiz master, b did not know whether the money was in the last box, but after the announcement, b knows that the last box is empty. Since the quiz master made her announcement on the basis b's knowledge of the game in the initial situation, b cannot, if he is a careful reasoner, assume that the announcement will remain to be true after b has been given new information about the game.

Consider now the possibility that models the situation before the announcement of the quiz master. For each $i \leq 5$, there is a possibility w_i that models the situation where box i contains the money. The information of b in all these possibilities is the same: he does not know in which box the money is; he only knows that it must be in some unique box. In a definition and a picture, we set, for each $i \leq 5$:

$$\begin{array}{rcl} w_i(p_i) & = & 1 \\ w_i(p_j) & = & 0 \text{ for } i \neq j \\ w_i(b) & = & \{w_1 \ldots w_5\} \end{array}$$

If the last box is filled, i.e. when the initial situation is w_5, then b will win the game. It holds that $w_5 \models \mathsf{win}$. If one of the middle boxes contains the money, then b will not win the game: when $i < 5$, then $w_i \models \neg\mathsf{win}$. Whatever the number i is, in the

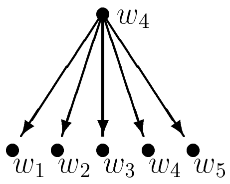


Figure 7.1. The initial situation with 5 boxes

world w_i, b does not know whether he will win or not: $w_i \models \neg\Box_b\mathsf{win} \wedge \neg\Box_b\neg\mathsf{win}$. All this is in accordance with our informal description of the 'paradox.'

In the initial situation w_4, the quiz master can indeed truthfully say that b cannot win. We assume that the effect of her announcement is that b consciously updates with the information expressed by $\neg\mathsf{win}$. From the announcement that $\neg\mathsf{win}$ is true, b learns that the situation must be one of the w_i with $i < 5$. Applying the definitions of DEL, the possibility w'_4 that results from updating w_4 with $U_b?\neg\mathsf{win}$ is given by the following equations for each $i \leq 4$:

$$\begin{array}{rcl} w'_i(p_j) & = & w_i(p_j) \text{ for each } j \\ w'_i(b) & = & \{w'_1 \ldots w'_4\} \end{array}$$

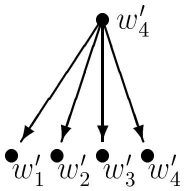


Figure 7.2. The possibility $w_4[\![U_b\neg\mathsf{win}]\!]$

In w'_4, the possibility that results after an update of w_4 with $U_b\neg\mathsf{win}$, b knows that the box 5 is empty. Since the money is in fact in box 4, b will win the game. It holds that $w'_4 \models \mathsf{win}$. So, although the quiz master spoke the truth when he said that b could not win in the initial situation, the effect of her announcement is that the situation changes to the effect that b will win the game after all.[5]

An extreme case of this situation is the case where there are only two boxes, and the money is in the first box. In this situation, it is true that b cannot win the game. But after b has learned that he cannot win, he can conclude that the money

[5]Consider also the following result. If n is the total number of boxes, i the number of the box that the money is in, and $[\pi]^n$ stands for a sequence of operators $[\pi]$ of length n, it holds that $w_i^n \models [U_b\neg\mathsf{win}]^{n-i-1}\neg\mathsf{win}$, while $w_i^n \models [U_b\neg\mathsf{win}]^{n-i}\mathsf{win}$. In other words, the present analysis predicts that if the full box is of distance $d > 0$ from the last box, the quiz master can say truthfully that b cannot win d times, but after the d-th time, b has received enough information to win the game after all. Another observation is that the only way that b can arrive at a state where he knows that he will win by updating a number of times with $\neg\mathsf{win}$ is when the money is in fact in the first box. All this under the assumption that b and the quiz master are perfect reasoners, and that their perfection is common knowledge between them.

was in the first box after all, and therefore that he will win. Here, we have a case in which b concludes, on the basis of the information that he cannot win, that in fact he will win. There is a whole lot of strangeness in this, but no contradiction. A sentence such as win, or Moore's paradox, or the negative answers in the puzzle of the dirty children, are such that the very act of uttering such a sentence can change the situation in such a way that its truth-value changes from truth to falsity.[6]

The analysis of the puzzle of the surprise exam I propose here can be summarized as follows. I have argued that the paradoxical result rests on a confusion between the interpretation of the quiz master's announcement as saying (1) "On the basis of what you know now, you cannot win the game," and (2) "On the basis of what you know now, or can infer from this very sentence, you cannot win the game." Sentence (2) is an example of a classical epistemic self-referential 'paradox,' which is interesting in its own right, but does not concern what I have taken to be the main problem of the surprise examination puzzle, namely that the quiz master says something that is plainly true, while it is contradictory to learn for the player. The analysis given here was mostly concentrated with the first reading of the announcement. I have tried to argue that dynamic epistemic semantics offers a natural analysis of how one can learn that sentences such as (1) are true while still not coming to believe them. One can even learn such a sentence and come to believe the contrary of the sentence. There is no contradiction involved here: the act of uttering such a sentence successfully may change the situation in such a way that the sentence become false.

7.3 Dialogue Games

In this section, I will define a simple dialogue game. The rules of the game are based on Grice's maxim's (Grice (1989)). I will treat these maxims as a kind of logical axioms of dialogue, to see what can be 'derived' from them. This is not in agreement with Grice's work at all; not in general because of his doubts about formalising pragmatic aspects, and not in particular, because the stress in his work is not on the situation where dialogue participants simply follow the maxims, but on situations where they do not. Instead, I will define a dialogue game that is mathematically very precise, in which the agents are assumed to always follow the rules. In a sense, the contents of this section are a non-Gricean formalisation of some aspects of Grice's work.

The game that will be defined in this section is not very sophisticated, and can only be seen as a first step in a richer theory. Still, I believe some aspects of it are interesting.

[6]It is perhaps interesting to compare such sentences with utterances such as "I am not saying something now," or performatives such as "I hereby pronounce you man and wife," the truth of which also depends on whether they are uttered or not.

We will assume that there exist only two agents, a and b. These two dialogue participants communicate in propositional modal logic, making 'utterances' of sentences. An 'utterance' is represented as $(s : \phi)$, where ϕ is a sentence, and s the agent making the utterance (the 's' stands for 'speaker'). I will use h as a meta-variable for the hearer, which, by convention, always denotes the agent different from s. I will assume that at each moment in the game the information of the agents can be represented by an information state, and hence, that the whole game-state can be represented by a possibility.

Definition 7.5 (utterances)
An *utterance* is a pair $(s : \phi)$, where ϕ is a sentence of multi-modal propositional logic $\mathcal{L}_{\{a,b\}}$, and $s \in \{a, b\}$ is an agent.
A *dialogue* is a sequence of utterances. □

In defining a dialogue game, we have to be precise about the rules of the game. A specification of a dialogue game involves at least the following two things:
(1) When a speaker is allowed to make an utterance. In other words, what the preconditions of an utterance $(s : \phi)$ in a situation w are.
(2) What the context-change potential of an utterance $(s : \phi)$ is, i.e. how a possibility w changes as the result of an utterance of ϕ by s.
These two are of course not independent. In the way I will set it up here, the context change potential of an utterance is directly based on preconditions of the utterance.

Grice (1989) has formulated a set of general rules that dialogue participants follow if they engage in a 'cooperative' dialogue. For the present game, I will define two preconditions for an utterance that loosely correspond to the first two Gricean maxims: the maxim of quality and that of quantity.

Grice's maxim of quantity is the following:

Quality

- Do not say what you believe to be false.
- Do not say that for which you lack adequate evidence.

Grice's maxim of quality is very similar to the condition of truthfulness of Lewis (1975).[7] Of course, I have said very little about 'evidence', let alone 'adequate evidence,' in this dissertation. I will simplify the maxim of quality by rephrasing it in terms of belief only.

Definition 7.6 (quality)
An utterance $s : \phi$ is qualitatively correct in a situation w iff

[7]He states on page 7: "My proposal is that the convention whereby a population P uses a language $\mathcal{L}$ is a convention of *truthfulness* and *trust* in $\mathcal{L}$. To be truthful in $\mathcal{L}$ is to act in a certain way: to try never to utter any sentences of $\mathcal{L}$ that are not true in $\mathcal{L}$. Thus it is to avoid uttering any sentence of $\mathcal{L}$ unless one believes it to be true in $\mathcal{L}$."

- $w \not\models \Box_s \neg\phi$ (the speaker does not believe that what she says is false)
- $w \models \Box_s \phi$ (the speaker believes what she says) □

We can shorten this definition by saying that an utterance of ϕ is qualitatively correct just in case the speaker is in a consistent state in which she believes that ϕ.

This simple rule is already productive. First of all, it rules out utterances of contradictions: those can never be correct, because they can never be consistently believed. More interestingly, the maxim of quality explains why Moore's paradox seems to be a paradox. This classical 'paradox' is the fact that although the sentence 'The cat is on the mat but John does not believe it' is perfectly consistent, an utterance of the sentence 'The cat is on the mat, but I, John, do not believe it' by John is never correct. In our simple framework, we can formalize the latter utterance as $(s : p \wedge \neg\Box_s p)$, where '$p$' stands for 'the cat is on the mat.' Of course, the sentence uttered is perfectly consistent. However, in an introspective possibility, such an utterance can never be qualitatively correct: if w is introspective, then $w \models \Box_s \neg(p \wedge \neg\Box_s p)$.

Another general principle of dialogue according to Grice is that utterances should be informative. Grice formulates the maxim as follows:

Quantity

1. Make your contribution as informative as required (for the current purposes of the exchange).
2. Do not make your contribution more informative than is required.

I will interpret this maxim in a very simple way, and shorten it to 'Be informative!'[8] But also this simplified imperative is open for further interpretation. One plausible reading is that to be informative, a sentence should provide the hearer with new information, but I will also consider a weaker form of the maxim of quantity, in which the information expressed by the sentence should not be common knowledge already.

If we assume that an utterance should always be informative for the hearer, there are several kinds of utterances that are predicted to be infelicitous. Firstly, it is predicted that utterances that are about the information of the hearer are not correct. Sentences such as "You know that the cat is on the mat" or "You don't know that the cat is on the mat" would be infelicitous, because they express information that, when true, is already known by the hearer (whose information is introspective). However, such sentences can be, and sometimes are, used to

[8] Groenendijk and Stokhof (1984) define a way of ordering sentences by how well they answer a given question. We can model the notion of a topic of a dialogue by a question (as is done, for example, in Kuppevelt (1991)). Combining the two ideas would give us a more sophisticated analysis of the maxim of quantity.

establish facts about the dialogue situation itself ("Having established your ignorance, I will proceed by enlightening you"). So adopting a maxim that completely rules out such sentences as incorrect may be too strong.

Consider also the case of the dirty children in this light. When the father says that at least one child is dirty, he is saying something that is already known by each of the children. If an utterance is felicitous only if the sentence is not yet believed by the audience, then his would be infelicitous. As we have seen, his utterance provides information anyway because it is not yet common knowledge that the sentence in question is true: the common knowledge between the children changes as the result of the utterance.

I am not sure whether the observation that the father's utterance is infelicitous under the strong interpretation of the maxim of quantity is a reason to adopt the weaker version or not. One could also argue that utterances about the information of the hearer (and, in general, utterances of sentences that are known to be believed by the hearer) are of a different category, and that the maxims do not apply to them: utterances about the information of speaker and hearer are not used for exchange of information proper, but to establish information about the dialogue situation itself, preparing the way, as it were, for further exchange about the 'real' topic of the conversation.

To be sure to err on the safe side, I will choose for the second, weaker condition. For an utterance to be quantitatively correct, the speaker must utter sentences that are not in the common ground. The condition that utterances should be informative with respect to the common ground is also proposed by Stalnaker (1978) (according to him, the principle 'can be defended as an essential condition of rational communication'), while Lewis (1975) proposes a version of this condition as a precondition of any "serious communication situation."[9]

Definition 7.7 (quantity)
An utterance (s, ϕ) is quantitatively correct in w iff the speaker s does not believe that it is common knowledge that ϕ is the case: $w \models \Diamond_s \neg C_{\{s,h\}}\phi$ □

Remember from section 3.2 that $w \models \Diamond_s \neg C_{\{s,h\}}\phi$ just in case $w \models \Box_s \neg C_{\{s,h\}}\phi$, if the information of s is consistent and introspective. That means that an agent is always fully informed about whether a certain sentence is common knowledge or not.

Information Change

We are now ready to turn to the second aspect of the dialogue game: that of specifying what the effect of an utterance is on a possibility. There are several options open for defining the context change potential of an utterance in a cooperative

[9] Lewis also requires that it is common knowledge that the speaker knows whether ϕ is true. I am not sure why he does.

dialogue. Suppose that there is no problem with the 'communication channel:' it is common knowledge between the participants that what they say is heard and understood by all participants. Suppose also that it is common knowledge that the speaker is following the maxims. In such a case, the *least* a hearer will learn from the fact that a speaker s makes an utterance of ϕ is that (1) s believes that ϕ is true and (2) s believes that ϕ is not common knowledge. But since the cooperativity and the trustworthiness of the communication channel are common knowledge, also the speaker will learn that the hearer will learn that the speaker believes both that ϕ as well as that ϕ is not common knowledge. In Lewis' terms, we assume that the agents are 'trusting,' and that the fact that they are trusting is common knowledge.[10]

In short, the minimal effect of an utterance $(s : \phi)$ in the game will be that speaker and hearer mutually learn that the speaker has uttered ϕ correctly.

Definition 7.8 (context change as the result of a correct utterance of $(s : \phi)$)

$$w[\![(s:\phi)]\!]v \quad \text{iff} \quad w[\![U_{\{s,h\}}?(\Box_s\phi \wedge \Box_s\neg C_{\{s,h\}}\phi)]\!]v \qquad \Box$$

The effect of an utterance $(s : \phi)$ is that s and h mutually learn that s believes that ϕ and that s believes that ϕ is not common knowledge. An update with $(s : \phi)$ not only changes the information state of the hearer, but also that of the speaker herself: she learns that the hearer has learned $(s : \phi)$ was uttered correctly.

Let us first consider the effect of $(s : \phi)$ on S5-possibilities, where agents cannot be mistaken (or disagree). First of all, in such models, it holds that $\Box_s\neg C_{\{s,h\}}\phi$ iff $C_{\{s,h\}}\neg C_{\{s,h\}}\phi$. In other words, if (s, ϕ) is a correct utterance in w, it will already be common knowledge in w that ϕ is not common knowledge. In such cases we can simplify the definition of an update with (s, ϕ) to being equivalent with $[\![U_{\{s,h\}}\Box_s\phi]\!]$.

Moreover, there is the fact that in S5-models, $\Box_s\phi$ implies that ϕ is the case. So, in these cases, an update with $[\![U_{\{s,h\}}\Box_s\phi]\!]$ is *stronger* than a mutual update with ϕ only. We can make this precise by using the information ordering of section 5.4.1.

Proposition 7.9 In S5-models w where (s, ϕ) is correct, it holds that:

$$w[\![(s,\phi)]\!] \text{ contains at least as much information as } w[\![U_{\{s,h\}}?\phi]\!] \qquad \Box$$

The extra information a hearer gets from $(s : \phi)$, over and above the information contained in ϕ itself, is that $(s : \phi)$ was a correct utterance to make. Somewhat more tendentiously, we could say that this extra information consists of the

[10] Lewis (1975) writes: "To be trusting in $\mathcal{L}$ is to form beliefs in a certain way: to impute truthfulness in $\mathcal{L}$ to others, and thus to tend to respond to another's utterance of any sentence of $\mathcal{L}$ by coming to believe that the uttered sentence is true in $\mathcal{L}$." See also Fagin and Halpern (1988).

implicatures of $(s : \phi)$. In the present case, these 'implicatures' consist only of the statements that the speaker believes ϕ, and believes ϕ not to be common knowledge.

The result of an utterance, in S5-models, will often, but not always, be that the sentence in question becomes common knowledge. This follows from proposition 4.17.

Proposition 7.10 (success of utterances)
If w is a S5-model in which $(s : \phi)$ is correct, and ϕ does not contain any negative occurrences of $\Box_s\phi$ or $\Box_h\phi$, then

$$w[\![(s : \phi)]\!] \models C_{\{s,h\}}\phi \qquad \Box$$

Exceptions to the success of an utterance are the puzzles of the previous sections.

In possibilities that are not reflexive, the hearer can, in general, not conclude that ϕ is the case from the fact that the speaker believes that ϕ. In such cases, the hearer can do one of two things. Either he believes the speaker with respect to ϕ, and thereby learns that ϕ is the case from the fact that the utterance $(s : \phi)$ is made, or she does not believe the speaker on this point, and refuses to draw the conclusion that ϕ is the case.

In both cases, a reaction on the part of the hearer is called for: either to the effect that he believes her with respect to her utterance of ϕ, or that he doesn't. The argument that, supposedly, will ensue between the dialogue participants if the hearer dissents falls beyond the scope of the simple framework presented here. I will concentrate on the first case instead.[11]

To introduce some further notation, let $(s : \phi; h : \text{ok})$ stand for an utterance by s of ϕ to which h has assented. The effect of $(h : \text{ok})$ after $(s : \phi)$ would be that s and h mutually learn that the hearer believes that the speaker is right with respect to ϕ. More formally, the effect would be that of $U_{\{s,h\}}?\Box_h(\Box_s\phi \to \phi)$. The effect of the whole utterance, together with the assent, would be this:

$$w[\![(s : \phi; h : \text{ok})]\!]v \quad \text{iff} \quad w[\![(s : \phi)]\!][\![U_{\{s,h\}}?\Box_h(\Box_s\phi \to \phi)]\!]v$$

Note that in S5-models, the update that represents the hearer's assent is completely vacuous.

In any case, one effect of an utterance of a simple indicative sentence will be that this sentence becomes common knowledge after a successful utterance:

Proposition 7.11 (success of acknowledged utterances)
If w is a K45-model, and ϕ is any sentence without negative occurrences of $\Box_s$ or $\Box_h$ that is correct in w, then:

$$w[\![(s : \phi, h : \text{ok})]\!] \models C_{\{s,h\}}\phi$$

[11] Clark and Schaefer (1989) contains a more empirically oriented discussion of 'acceptance moves' in dialogue.

□

This proposition shows that in general a successful utterance of an indicative sentence will make it common knowledge that this sentence is true.

The fact that an utterance is correct has some pleasant consequences. Firstly, it holds that an update with a correct utterance preserves reflexivity and introspection. This means that if the agents start out in a situation where their information is factual, they will never reach a situation in which their information is false by uttering correct sentences.

Proposition 7.12 If w is reflexive and introspective, and $(s : \phi)$ is correct in w, then $w[\![(s : \phi)]\!]$ is reflexive and introspective. □

Secondly, it holds that correct utterances change the information states of the dialogue participants without changing the combined information of the dialogue participants that was defined in section 3.4. If two agents are exchanging information and do not get any new information from the outside, then it should never happen that their combined information grows during the exchange. This is indeed the case:

Proposition 7.13
If $(s : \phi)$ is correct in w, and w is correct, then the combined information in w is the same as that in $w[\![(s : \phi)]\!]$ □

This proposition expresses that the combined information of speaker and hearer can never grow when they make correct utterances, and when utterances are interpreted in the way I sketched above. In other words, they cannot get information from each other that is not already present in their information states 'taken together.' The condition that the combined information of agents never grow in a correct dialogue is an extension of Dekker (1993)'s notion of *proper information exchange* to the higher-order case.

Legal dialogues

Having abruptly concluded the discussion of the context change potential of utterances in a belief model, I now turn to the question what kind of dialogues are possible given the two simple rules above.

Definition 7.14

- A *dialogue* is a sequence of utterances.
- A dialogue $(a_1 : \phi_1); \ldots ; (a_n : \phi_n)$ is *correct* in w iff $(a_{i+1} : \phi_{i+1})$ is (qualitatively and quantitatively) correct in $w[\![(a_1 : \phi_1)]\!] \ldots [\![(a_i : \phi_i)]\!]$ for each $i < n$.

  A dialogue is *legal* iff there is a reflexive and introspective possibility in which it is correct. □

The term 'legal dialogue' is from Hamblin (1971). It is similar to the notion of a coherent text of Groenendijk et al. (1996).

A legal dialogue is simply a sequence of utterances for which we can find a context (a possibility) where the first utterance is correct, and where the second utterance is correct after 'processing' the first utterance, etcetera. The condition that this context be one that is introspective and reflexive guarantees that we are dealing with a 'normal' context: a dialogue is correct just in case we can find a context for that dialogue in which speaker and hearer trust each other completely.

We are using a very simple notion of correctness here: the only restrictions are that the speaker is in a consistent state in which she believes that the sentence she utters is true (quality), and that this sentence is not yet common knowledge (quantity).

If we look at dialogues in which only sentences of propositional logic are uttered, it can be proven that a dialogue is legal exactly when it does not contain a contradiction or a repetition (in the sense that utterances should never contradict previously given information, nor should they follow from information that is already given).

Proposition 7.15 Let $(a_i : \phi_i)_{i\leq n}$ be a dialogue such that each ϕ_i is propositional. It holds that:

$$
\begin{array}{ll}
(a_i : \phi_i)_{i\leq n} \text{ is legal} & \text{iff} \\
& \text{(1) There is no } j \leq n \text{ such that } \bigwedge_{i<j} \phi_i \models \neg\phi_j \\
& \text{(2) There is no } j \leq n \text{ such that } \bigwedge_{i<j} \phi_i \models \phi_j
\end{array}
$$

proof: We prove by induction on the length n of the dialogue that (1) and (2) hold. Assume we have a dialogue $(a_i : \phi_i)_{i\leq n+1}$ that is legal. Then, there must be an introspective and reflexive w in which the dialogue is correct. By definition of legality, also $(a_i : \phi_i)_{i\leq n}$ must be legal. By induction hypothesis, we may assume that (1) and (2) are true for $j \leq n$, and we only need to show (1) and (2) in the case where $j = n + 1$.

Let v be the result of updating w with the dialogue up to utterance n, i.e. $w[\![(a_i : \phi_i)_{i\leq n}]\!]v$. Since each ϕ_i is propositional, and propositional sentences are persistent over updates, we can conclude with proposition 7.10 that $v \models C_{\mathcal{A}} \bigwedge_{i\leq n} \phi_i$. Since $(a_{n+1} : \phi_{n+1})$ is qualitatively correct, it must hold that $v(a_{n+1})$ is consistent, and that $v \models \Box_{a_{n+1}}\phi$. Since also $v \models \Box_a \bigwedge_{i\leq n} \phi_i$, it follows that $\bigwedge_{i\leq n} \phi_i \not\models \neg\phi_{n+1}$, which takes care of (1). Since $(a_{n+1} : \phi_{n+1})$ must be qualitatively correct as well, it must hold that $v \models \Box_{a_{n+1}}\neg C_{\mathcal{A}}\phi_{n+1}$. Since $v \models \Box_{a_{n+1}} C_{\mathcal{A}} \bigwedge_{i\leq n} \phi_i$, it holds that $\bigwedge_{i\leq n} \phi_i \not\models \phi_{n+1}$.

For the other direction, assume that (1) and (2) hold. Then $\bigwedge_{i\leq n} \phi_i$ is consistent. Consider now a state w in which all dialogue participants believe that all sentences in the dialogue are true, but they also all think it is possible that one of the others does not know of any of the sentences in that dialogue that they are true. It is not hard to check (by induction on the length of the dialogue) that the dialogue must be correct in this state. $\square$

A dialogue is legal iff each sentence uttered in the dialogue is not already implied by the previous sentences, nor contradicts any of the previous sentences. The set of legal propositional dialogues are exactly the set of legal dialogues in Hamblin (1971)'s 'System 2.'

It is important to see that proposition 7.15 above cannot be generalized to all dialogues. Dialogues that contain utterances of sentences that are about the information of the dialogue participants themselves can be legal, yet contain repetitions or even truth-conditional contradictions. For example, the dialogue $(a : \Diamond_a \phi); (b : \neg\phi); (a : \Box_a \neg\phi)$ is legal, although it does contain utterances of sentences that (truth-conditionally) contradict each other. The puzzle of the dirty children shows that there are legal dialogues that contain repetitions.

Since at each point in a dialogue, a correct utterance must bring in something new to the dialogue, without contradicting the previous, it may happen that at a certain moment, nothing can be correctly uttered anymore. In that case, we say that the dialogue is finished.

Definition 7.16 A dialogue Ψ is finished from w iff there is no utterance that is correct in $w[\Psi]$. $\square$

A dialogue is finished in a possibility w iff nothing can be correctly said anymore after the dialogue in w. A dialogue is finished iff each agent thinks that everything he believes is common knowledge already.

Proposition 7.17 A dialogue Φ is finished iff $w[\Phi] \models \Box_a \phi \rightarrow \Box_a C_{\{a,b\}} \phi$ for all ϕ. $\square$

A typical case of a possibility in which nothing can correctly be said is one in which the state of each of the dialogue participants is equal to the state that represents their combined information. When a speaker believes that everything he knows is common knowledge, he cannot utter anything correctly anymore.

Bibliography

Peter Aczel (1988) *Non-Well-Founded Sets.* CSLI Lecture Notes 14. CSLI Publications, Stanford.

Carlos E. Alchourrón, Peter Gärdenfors and David Makinson (1985) *On the Logic of Theory Change: Partial Meet Contractions and Revision Functions.* Journal of Symbolic Logic, 50:510–530.

Robert J. Aumann (1976) *Agreeing to disagree.* Annals of Statistics, 4(6):1236–1239.

Alexandru Baltag, Lawrence S. Moss and Slawomir Solecki (to appear) *The Logic of Public Announcements and Common Knowledge.* To appear in the Proceedings of the Conference on Theoretical Aspects of Reasoning about Knowledge 1998. `http://www.math.indiana.ed/home/moss/articles.html`.

Jon Barwise (1981) *Scenes and Other Situations.* The Journal of Philosophy, 78(1):369–397.

Jon Barwise (1989) *On the Model Theory of Common Knowledge.* In: The Situation in Logic, CSLI Lecture notes 17, pages 201–220. CSLI Publications, Stanford.

Jon Barwise and John Etchemendy (1987) *The Liar: An Essay on Truth and Circularity.* Oxford University Press, New York.

Jon Barwise and Lawrence S. Moss (1996) *Vicious Circles.* CSLI Publications, Stanford.

David Beaver (1995) *Presupposition and Assertion in Dynamic Semantics.* Center for Cognitive Science, University of Edinburgh.

Nuel D. Belnap (1977) *A Useful Four-Valued Logic.* In: J. Michael Dunn and George Epstein, eds. (1977) Modern Uses of Multiple-Valued Logic. Reidel, Dordrecht.

Johan van Benthem (1976) *Modal Correspondence Theory.* Mathematisch Instituut en Instituut voor Grondslagen Onderzoek, Universiteit van Amsterdam.

Johan van Benthem (1991) *Language in Action. Categories, Lambdas and Dynamic Logic.* Volume 130 of Studies in Logic. North-Holland, Amsterdam.

Johan van Benthem (1996) *Exploring Logical Dynamics.* Studies in Logic, Language and Information. CSLI Publications, Stanford.

Anton Benz and Gerhard Jäger, eds. (1997) *Proceedings of the Munich Workshop on Formal Semantics and Pragmatics of Dialogue.* CIS, München.

Tijn Borghuis (1994) *Coming to Terms with Modal Logic: on the Interpretation of Modalities in Tsyped λ-Calculus.* Technische Universiteit Eindhoven.

Herbert H. Clark and Catherine R. Marshall (1981) *Definite Reference and Mutual Knowledge.* In: A.K. Joshi, B. L. Webber and I. A. Sag, eds. (1981) Elements of Discourse Understanding, pages 10–63. Cambridge University Press, Cambridge.

Herbert H. Clark and E. Schaefer (1989) *Contributing to Discourse.* Cognitive Science, 13(259–294).

Paul Dekker (1993) *Transsentential Meditations: Ups and Downs in Dynamic Semantics.* ILLC Dissertation Series 1993-1, University of Amsterdam.

Keith Devlin (1992) *The Joy of Sets: Fundamentals of Contemporary Set Theory.* Undergraduate Texts in Mathematics. Springer Verlag, Berlin.

Kees Doets (1987) *Completeness and Definability: Applications of the Ehrenfeucht Game in Second-Order and Intensional Logic.* Stichting Mathematisch Centrum, Amsterdam.

Jan van Eijck and Fer-Jan de Vries (1995) *Reasoning about Update Logic.* Journal of Philosophical Logic, 24(2):19–45.

Peter van Emde Boas, Jeroen Groenendijk and Martin Stokhof (1980) *The Conway Paradox: Its Solution in an Epistemic Framework.* Proceedings of the Third Amsterdam Montague Symposion, pages 159–183.

Ronald Fagin (1994) *A Quantitative Analysis of Modal Logic.* The Journal of Symbolic Logic, 59(1):209–252.

Ronald Fagin and Joseph Y. Halpern (1988) *I'm OK if You're OK: On the Notion of Trusting Communication.* Journal of Philosophical Logic, 17(1):329–354.

Ronald Fagin, Joseph Y. Halpern, Yoram Moses and Moshe Vardi (1995) *Reasoning about Knowledge.* The MIT Press, Cambridge, Massachusetts.

Ronald Fagin, Joseph Y. Halpern and Moshe Y. Vardi (1991) *A Model-Theoretic Analysis of Knowledge.* Journal of the Association for Computing Machinery, 38(2):382–428.

Ronald Fagin, Joseph Y. Halpern and Moshe Y. Vardi (1992) *What Can Machines Know? On the Properties of Knowledge in Distributed Systems.* Journal of the Association for Computing Machinery, 39(2):328–376.

Ronald Fagin and Moshe Y. Vardi (1985) *An Internal Semantics for Modal Logic: Preliminary Report.* In: Proceedings of the Seventeenth Annual ACM Symposium on Theory of Computing, pages 305–315. ACM, New York.

George Gargov and Solomon Passy (1990) *A Note on Boolean Modal Logic.* In: Petio Petrov Petkov, ed. (1990) Mathematical Logic, Proceedings of the Summer School and Conference on Mathematical Logic, honourably dedicated to the ninetieth anniversary of Arend Heyting (1889–1980) held September 13–23, 1988, in Chaika (near Varna), Bulgaria, pages 299–309. Plenum Press.

Jelle Gerbrandy (1994) *'Might' in Update Semantics.* Master's thesis, University of Amsterdam.

Jelle Gerbrandy (1997a) *Bisimulation and Bounded Bisimulation.* ILLC Research Report LP–1997–05.

Jelle Gerbrandy (1997b) *Changing the Common Ground.* In: Anton Benz and Gerhard Jäger, eds. (1997) Proceedings of the Munich Workshop on Formal Semantics and Pragmatics of Dialogue, pages 40–58. CIS, München.

Jelle Gerbrandy (1997c) *Dynamic Epistemic Logic.* ILLC Research Report LP–1997–04. To appear in the Proceedings of the Second Conference on Information-Theoretic Approaches to Logic, Language, and Computation.

Jelle Gerbrandy (1998) *Distributed Knowledge.* In: Joris Hulstijn and Anton Nijholt, eds. (1998) Twendial'98: Formal Semantics and Pragmatics of Dialogue, TWLT 13, pages 111–124. Universiteit Twente, Enschede.

Jelle Gerbrandy and Willem Groeneveld (1997) *Reasoning about Information Change.* Journal of Logic, Language, and Information, 6:147–169.

Robert Goldblatt (1987) *Logics of Time and Computation.* CSLI Lecture Notes 7. CSLI Publications, Stanford.

Paul Grice (1989) *Logic and Conversation.* In: Studies in the Way of Words, pages 22–40. Harvard University Press, Cambride, Massachusetts.

Jeroen Groenendijk and Martin Stokhof (1984) *Studies on the Semantics of Questions and the Pragmatics of Answers.* Jurriaans BV, Amsterdam.

Jeroen Groenendijk and Martin Stokhof (1991) *Dynamic predicate logic.* Linguistics and Philosophy, 14(1):39–100.

Jeroen Groenendijk and Martin Stokhof (1997) *Questions.* In: Johan van Benthem and Alice ter Meulen, eds. (1997) Handbook of Logic and Language, pages 1055–1124. Elsevier.

Jeroen Groenendijk, Martin Stokhof and Frank Veltman (1996) *Coreference and Modality.* In: Shalom Lappin, ed. (1996) Handbook of Contemporary Semantic Theory, pages 179–213. Blackwell, Oxford.

Willem Groeneveld (1995) *Logical Investigations into Dynamic Semantics.* ILLC Dissertation Series 1995-18.

Joseph Y. Halpern (1987) *Using Reasoning about Knowledge to Analyze Distributed Systems.* In: J.F. Traub, B.J. Grosz, B.W. Lampson and N.J. Nilsson, eds. (1987) Annual Review of Computer Science, Volume 2, pages 37–68. Annual Reviews Inc., Palo Alto, California.

Joseph Y. Halpern and Yoram Moses (1990) *Knowledge and Common Knowledge in a Distributed Environment.* Journal of the Association for Computing Machinery, 37(3):549–587.

Charles Leonard Hamblin (1971) *Mathematical models of dialogue.* Theoria, 2:130–155.

Sharon J. Hamilton and James P. Delgrande (1989) *An Investigation of Modal Structures as an Alternative Semantic Basis for Epistemic Logics.* Computational Intelligence, 5(2):82–96.

David Harel (1984) *Dynamic Logic.* In: Dov M. Gabbay and Franz Guenthner, eds. (1984) Handbook of Philosophical Logic, Vol. 2, pages 497–604. Reidel, Dordrecht.

Irene Heim (1982) *The Semantics of Definite and Indefinite Noun Phrases.* University of Massachusetts, Amherst. Published in 1989 by Garland, New York.

Matthew Hennessy and Robin Milner (1985) *Algebraic Laws for Nondeterminism and Concurrency.* Journal of the Association for Computing Machinery, 32(1):137–161.

Wolfgang Heydrich and Hannes Rieser, eds. (1998) *Mutual Knowledge, Common Ground and Public Information, Workshop at the 10th European Summer School in Logic, Languages and Information* (1998).

Risto Hilpinen (1969) *An Analysis of Relativised Modalities.* In: J.W. Davis, D.J. Hockney and W.K. Wilson, eds. (1969) Philosophical Logic, pages 181–193.

Risto Hilpinen (1974) *On the Semantics of Personal Directives.* In: Carl H. Heydrich, ed. (1974) Semantics and Communication, pages 162–179. North-Holland Publishing Co., Amsterdam.

Jaakko Hintikka (1962) *Knowledge and Belief.* Cornell University Press.

Wiebe van der Hoek and John-Jules Ch. Meyer (1992) *Making Some Issues of Implicit Knowledge Explicit.* International Journal of Foundations of Computer Science, 3(2):193–223.

Wiebe van der Hoek and John-Jules Ch. Meyer (1997) *A Complete Epistemic Logic for Multiple Agents: Combining Distributed and Common Knowledge.* In: Michael Bacharach, Z. A. Gerbrand-Varet, P. Mongin and H. S. Shin, eds. (1997) Epistemic Logic and the theory of Games and Decisions, pages 35–68.

Wiebe van der Hoek, Bernd van Linder and John-Jules Ch. Meyer (1994a) *Communicating Rational Agents.* In: B. Nebel and L. Dreschler-Fische, eds. (1994) KI-94: Advances in Artificial Intelligence, Volume 861 of *Lecture Notes in Computer Science (subseries LNAI)*, pages 202–213. Springer Verlag, Berlin.

Wiebe van der Hoek, Bernd van Linder and John-Jules Ch. Meyer (1994b) *Tests as Epistemic Updates.* In: A. Cohn, ed. (1994) Proceedings of the 11th European Conference on Artificial Intelligence (ECAI'94), pages 331–335. John Wiley & Sons.

Wiebe van der Hoek, Bernd van Linder and John-Jules Ch. Meyer (to appear) *Group Knowledge Isn't Always Distributed.* Mathematics for the Social Sciences.

Wiebe van der Hoek, Bernd van Linder and John-Jules Ch. Meyer (unpublished). *An Integrated Modal Approach to Rational Agents.* Unpublished manuscript.

Marco Hollenberg (1995) *Hennessy-Milner Classes and Process Algebra.* In: Alban Ponse, Maarten de Rijke and Yde Venema, eds. (1995) Modal Logic and Process Algebra: a Bisimulation Perspective, CSLI Lecture Notes 53, pages 187–216. CSLI Publications, Stanford.

I. L. Humberstone (1985) *The Formalities of Collective Omniscience.* Philosophical Studies, 48:401–423.

Jan Jaspars (1994) *Calculi for Constructive Communication.* ILLC Dissertation

Series 1994-4, ITK Dissertation Series 1994-1, Katholieke Universiteit Brabant.

Hans Kamp and Uwe Reyle (1993) *From Discourse to Logic.* Kluwer Academic Publishers, Dordrecht.

Dexter Kozen and Rohit Parikh (1981) *An Elementary Proof of the Completeness of PDL.* Theoretical Computer Science, 14:113–118.

Saul Kripke (1963a) *A Semantical Analysis of Modal Logic I, Normal Propositional Calculi.* Zeitschrift für mathematische Logik und Grundlagen der Mathematik, 9:63–96.

Saul Kripke (1963b) *Semantical Considerations on Modal Logic.* Acta Philosophica Fennica, pages 83–94.

Jan van Kuppevelt (1991) *Topic en Comment: Expliciete en impliciete vraagstelling in discourse.* Katholieke Universiteit Nijmegen.

Fred Landman (1986) *Towards a Theory of Information.* Universiteit van Amsterdam. Also appeared as GRASS 6 with Foris Publications, Dordrecht, Holland/Cinnaminson, U.S.A.

Wolfgang Lenzen (1978) *Recent Work in Epistemic Logic.* Acta Philosphica Fennica, 30(1):1–219.

David Lewis (1969) *Convention, a Philosophical Study.* Harvard University Press, Cambridge, Massachusetts.

David Lewis (1975) *Languages and Language.* In: Keith Gunderson, ed. (1975) Minnesota Studies in the Philosophy of Science, Volume VII, pages 3–35. University of Minnesota Press, Minneapolis.

David Lewis (1979) *Scorekeeping in a Language Game.* Journal of Philosophical Logic, 8:339–359.

John McCarthy (1990) *Formalization of Two Puzzles Involving Knowedge.* In: Vladimir Lifschitz, ed. (1990) Formalizing Common Sense: Papers by John McCarthy, pages 1–61. Ablex.

Yoram Moses, Danny Dolev and Joseph Y. Halpern (1986) *Cheating Husbands and Other Stories: A Case Study of Knowledge, Action, and Communication.* Distributed Computing, 1:167–176.

Mogens Nielsen and Christian Clausen (1994) *Bisimulation for Models in Concurrency.* In: B. Jonsson and J. Parrow, eds. (1994) Concur '94: Concurrency Theory, Lecture Notes in Computer Science 836, pages 385–400. Springer Verlag, Berlin.

D. J. O'Connor (1948) *Pragmatic Paradoxes.* Mind, 57:358–359.

Rohit Parikh (1992) *Finite and Infinite Dialogues.* In: Y.N. Moschovakis, ed. (1992) Logic from Computer Science, Mathematical Sciences Research Institute publications 27, pages 481–497. Springer Verlag, Berlin.

David Park (1981) *Concurrency and Automata on Infinite Sequences.* In: P. Deussen, ed. (1981) Proceedings of the 5th GI Conference, Lecture Notes in Computer Science, Volume 104, pages 167–183. Springer Verlag, Berlin.

Solomon Passy and Tinko Tinchev (1991) *An Essay in Combinatory Modal Logic.*

Information and Computation, 93:263–332.

Vaughn R. Pratt (1976) *Semantical considerations on Floyd-Hoare logic.* Proceedings of the 17th IEEE Symposion on Foundations of Computer science, pages 109–121.

Willard Van Orman Quine (1953) *On a So-Called Paradox.* Mind, 62:65–67. Reprinted as 'On a Supposed Antinomy' in Quine (1966).

Willard Van Orman Quine (1966) *The Ways of Paradox and Other Essays.* Random House.

Maarten de Rijke (1993) *Extending Modal Logic.* ILLC Dissertation Series 1993-4, University of Amsterdam.

Maarten de Rijke (1994) *Meeting Some Neighbours.* In: Jan van Eijck and Albert Visser, eds. (1994) Logic and Information Flow. The MIT Press, Cambridge, Massachusetts.

Eric Rosen (1995) *Modal Logic over Finite Structures.* ILLC Research Report ML-95-08. `ftp://ftp.cis.upenn.edu/pub/ircs/tr/95-27.ps.Z`.

Stephen R. Schiffer (1972) *Meaning.* Clarendon Press, Oxford.

R. Shaw (1958) *The Paradox of the Unexpected Examination.* Mind, 67:382–384.

Robert C. Stalnaker (1972) *Pragmatics.* In: Harman and Davidson, eds. (1972) Semantics of Natural Language, pages 380–397. Reidel, Dordrecht.

Robert C. Stalnaker (1978) *Assertion.* In: Peter Cole, ed. (1978) Pragmatics (Syntax and Semantics 9), pages 315–332. Academic Press, New York.

Colin Stirling (1996) *Games and the Modal μ-Calculus.* In: Lecture Notes in Computer Science, 1055, pages 298–312. Springer Verlag, Berlin.

Frank Veltman (1996) *Defaults in Update Semantics.* Journal of Philosophical Logic, 25:221–261.

Michael Wooldridge and Nicholas R. Jennings (1994) *Agent Theories, Architectures, and Languages: A Survey.* In: Michael J. Wooldridge and Nicholas R. Jennings, eds. (1994) Intelligent Agents: Proceedings ECAI-94 workshop on Agent Theories, Architectures and Languages, pages 1–39. Springer Verlag, Berlin.

Henk Zeevat (1997) *The Common Ground as a Dialogue Parameter.* In Benz and Jäger (1997), pages 195–214.

Many of the ILLC Research Reports can be downloaded at:
`http://www.wins.uva.nl/research/illc/wwwreports.html`

Summary

Summarized in four words, the topic of this dissertation is Multi-Agent Dynamic Epistemic Semantics. The words 'semantics' implies that model theory plays a central role, the phrase 'epistemic' marks the fact that we are concerned with information (knowledge, belief), 'dynamic' stands for the fact that change of information is addressed, and the phrase 'multi-agent' says that there may be more than one agent involved.

Epistemic semantics has been a sub ject that can stand on its own since the work of Hintikka (1962) about knowledge and belief and the proposal of Kripke (1963b) for a semantics of modal logic. These authors were concerned with extending propositional logic with a modal operator (I will write '$\Box$'). Given a sentence ϕ, we can make a new sentence $\Box\phi$ that, in epistemic logic, is to be read as 'it is believed that ϕ.' The insight of Kripke and Hintikka was that this operator can be given a semantics using 'possibilities' and a relation of accessability between these possibilities. The accessability relation obtains between possibilities w and v exactly when v is compatible with what is believed in w. One can then define that a sentence of the form $\Box\phi$ is true in a world w just in case ϕ is true in all models that are compatible with what is known in w.

In this dissertation, this semantics is taken as a starting point. I study and develop extensions of this semantics in two dimensions: it is made 'multi-agent' and 'dynamic.'

The 'multi-agent' part is the topic of the third chapter. Here, I study an extension of epistemic logic where there is not one single modal operator $\Box$, but a whole family of operators of the form $\Box_a$ present in the language, where a is the name of an agent. A sentence $\Box_a\phi$ can be read as: 'Agent a believes that ϕ is true;' its semantics is just like that of $\Box$, except that for each agent a there is a corresponding accessability relation in the model. The semantics for this multi-agent logic makes it possible to define a number of new operators that are interesting from a philosophical as well as from a logical perspective. Already familiar from the literature are operators for 'common knowledge' and for 'distributed knowledge.' Both of these operators raise issues about the status of the semantics. Both have interesting completeness proofs as well. Newly introduced in this thesis is the concept of combined knowledge, which is very similar to, but different from, distributed knowledge. Moreover, I briefly discuss how the meaning of these operators can be made dependant on the 'topic' they are about.

The two chapters after that, chapters 4 and 5, are about the extension of epistemic logic with epistemic actions. I add 'programs' to the language of epistemic logic that denote changes in the world. For example, there is a program that expresses that the agents a and b both learn that c learnt that ϕ is true. In chapter 4, I define a semantics for this extension of the language, and give a sound and complete axiomatization of the resulting logic. The logic as a number of interesting properties. One of these is that learning that a sentence is true does not always mean that you reach a state in which you believe that this sentence is true. In the last chapter, this property of the logic is used to explain certain puzzling phenomena that arise in communication. Several authors have written about information change in epistemic logic before. In chapter 5, I compare the work of some of them with the semantics I defined in chapter 4.

In the last two chapters I show some ways that the semantics developed in this thesis can be useful for a formal analysis of dialogue. The sixth chapter is a discussion of some aspects of the concept of 'common ground' as it is used in theory of dialogue. In particular, it is about the relation between changes in the common ground and changes in the separate information states of the dialogue participants separately.

In the last chapter, I discuss two puzzles that are related to information change: the 'puzzle of the dirty children' and the 'surprise exam paradox.' Both of these puzzles can be straightforwardly described in the language that was defined in the earlier chapters. The associated effects of change of information provide an explanation of some of the more salient aspects of these puzzles. The dissertation ends with a description of a simple dialogue game.

I have not mentioned the first two chapters in this abstract yet. These chapters are more technical in character and are supposed to provide the background to the formal methods used in the thesis. Instead of the usual set-theory, I have used non-well-founded set-theory in modeling epistemic logic. Since this theory is not very familiar, I have thought it useful to add a general introduction to this topic. The main reason for using such a non-standard framework is that it allows for a more straightforward and elegant way of defining operations that change models.

Non-well-founded models for epistemic logic differ from the more standard Kripke models in that the latter way of modeling information allows for making distinctions that cannot be made in the former. In effect, using non-well-founded models means collapsing distinctions between Kripke models that are bisimilar. The second chapter contains a number of results about bisimulation and its relation with modal logic that together provide circumstantial evidence that collapsing distinctions between bisimilar models is harmless.

Samenvatting

Het onderwerp van dit proefschrift kan in zes woorden worden samengevat als 'Dynamische Epistemische Semantiek voor meerdere actoren.' Het woord 'semantiek' betekent dat model-theorie een centrale rol speelt, de frase 'epistemisch' geeft aan dat we geïnteresserd zijn in informatie (kennis, geloof), 'dynamisch' staat voor het feit dat verandering van informatie wordt besproken, en de toevoeging 'voor meerdere actoren' spreekt voor zich.

Epistemische semantiek is een onderwerp dat op eigen benen kan staan sinds het werk van Hintikka (1962) over kennis en geloof en het voorstel van Kripke (1963b) voor een semantiek van modale logica. Zij hielden zich bezig met de uitbreiding van propositionele logica met een modale operator (waarvoor ik het symbool '$\Box$' gebruik). Gegeven een zin ϕ kunnen we een nieuwe zin $\Box\phi$ maken die, in epistemische logica, gelezen moet worden als 'men gelooft dat ϕ waar is.' Het inzicht van Kripke en Hintikka was dat een semantiek voor deze operator gegeven kan worden met behulp van 'mogelijkheden' en een toegankelijkheidsrelatie tussen deze mogelijkheden.

In dit proefschrift wordt deze semantiek als uitgangspunt genomen. Ik bestudeer en ontwikkel uitbreidingen van deze semantiek in twee richtingen: het wordt 'multi-actor' en 'dynamisch' gemaakt.

In het derde hoofdstuk bestudeer ik een extensie van epistemische logica waarbij er niet één operator $\Box$, maar meerdere operatoren van de vorm $\Box_a$ aanwezig zijn in de taal, één voor iedere actor a. Een zin van de vorm $\Box_a\phi$ betekent 'actor a gelooft dat ϕ waar is.' De semantiek is dezelfde als die van $\Box$, behalve dat er nu met iedere actor a een eigen toegankelijksheidsrelatie in het model correspondeert. Deze semantiek maakt het mogelijk om een aantal nieuwe epistemische operatoren te definiëren die zowel vanuit een filosofisch als vanuit een logisch oogpunt interessant zijn. Operatoren voor gezamenlijke kennis en gedistribueerde kennis zijn al bekend uit de literatuur. De definities van deze operatoren roepen vragen over de status van de modellen op; ook hebben de logica's hebben interessante volledigheidsbewijzen. Nieuw in dit proefschrift is het concept van 'gecombineerde kennis' dat erg lijkt op gedistribueerde kennis, maar daar niet mee samenvalt. Ook bespreek ik in het kort hoe de betekenis van deze operatoren afhankelijk kan worden gemaakt van het onderwerp waar ze over gaan.

De twee daaropvolgende hoofdstukken, 4 en 5, gaan over de uitbreiding van epistemische logica met epistemische akties. Ik voeg 'programma's' toe aan de taal, die veranderingen in de informatie van de actoren uitdrukken. Zo is er een programma dat uitdrukt dat a en b samen leren dat c leert dat ϕ waar is. In het vierde hoofdstuk definieer ik een semantiek voor deze uitbreiding van de taal, en formuleer ik een volledig axiomasysteem voor de logica die daaruit voortvloeit. Deze logica heeft een aantal interessante eigenschappen. Eén daarvan is dat het leren dat een zin waar is niet altijd hetzelfde is als in een toestand terecht komen waarin ook geloofd wordt dat die zin waar is. In het vijfde hoofdstuk bespreek ik het werk van een aantal auteurs die over het onderwerp van informatieverandering hebben gepubliceerd, en vergelijk ik hun werk met de semantiek die ik in het vierde hoofdstuk gedefinieerd heb.

In de laatste twee hoofdstukken probeer ik te laten zien hoe de semantiek die in dit proefschrift ontwikkeld is nuttig kan zijn voor een formele analyse van dialogen. In het zesde hoofdstuk worden een aantal aspecten van het concept 'common ground,' zoals het in de literatuur over dialogen gebruikt wordt, besproken. Om preciezer te zijn gaat het hoofdstuk over de relatie tussen veranderingen in de common ground en de informatie van de verschillende betrokkenen.

In het laatste hoofdstuk bespreek ik twee puzzels die te maken hebben met informatieverandering: de puzzel van de vieze kinderen en de paradox van het onverwachte examen. Deze puzzels kunnen beide op een direkte en simpele manier beschreven worden in de taal die in dit proefschrift gedefinieerd is. De observatie dat het leren van een zin niet altijd betekent dat die zin ook als waar wordt beschouwd, speelt een belangrijke rol in deze uitleg. Het hoofdstuk eindigt met een beschrijving van een simpel dialoogspel.

Ik heb de eerste twee hoofdstukken van het proefschrift nog niet genoemd. Deze hoofdstukken zijn meer technisch van karakter. Deze hoofdstukken bevatten de theoretische achtergrond voor de formele methoden die in dit proefschrift worden gebruikt. In plaats van de standaardtheorie van verzamelingen heb ik gekozen voor het gebruik van niet-wel-gefundeerde verzamelingentheorie om epistemische toestanden te modelleren. Omdat deze theorie niet zo bekend is, leek het me nuttig het proefschrift te beginnen met een algemene introductie tot deze theorie.

Niet-wel-gefundeerde modellen voor epistemische logica verschillen van de meer standaard Kripkemodellen op het punt dat het bij de laatste mogelijk is bepaalde onderscheiden te maken die niet te maken zijn bij de eerste. Het komt erop neer dat bij het gebruik van niet-wel-gefundeerde modellen de verschillen tussen bisimulaire modellen wegvallen. Het tweede hoofdstuk bestaat uit een verzameling resultaten over bisimulatie en haar relatie met epistemische logica die het aanemelijk maken dat het wegvallen van het onderscheid tussen bisimulaire modellen onschuldig is.

Gearfetting

It ûnderwerp fan dit proefskrift kin yn seis wurden gearfette wurde as: 'Dynamyske Epistemyske Semantyk foar ferskate aktoaren'. It wurd 'semantyk' tsjut oan dat modelteory in sintrale rol spilet; de fraze 'epistemysk' jout oan dat wy te krijen hawwe mei ynformaasje (witten, leauwe); 'dynamysk' wiist derop dat feroaring fan ynformaasje besprutsen wurdt, en it taheakke 'foar ferskate aktoaren' wol sizze dat der mear as ien persoan yn meispilet.

Epistemyske semantyk is in ûnderwerp dat op himsels stean kin sûnt it wurk fan Hintikka (1962) oer witten en leauwen en it útstel fan Kripke (1963) foar in semantyk fan modale logika. Hja hâlden har dwaande mei it útwreidzjen fan de proposysjonele logika mei in modale operator (dêr't ik it symboal $\Box$ foar brûk). As ϕ in sin is, dan kinne wy in nije sin $\Box\phi$ meitsje dy't, yn epistemyske logika, lêzen wurde moat as 'der wurdt leaud dat ϕ wier is'. It ynsjoch fan Kripke en Hintikka wie dat jo foar dy operator in semantyk jaan kinne mei help fan in samling 'mooglikheden' en in tagonklikheidrelaasje tusken dy mooglikheden.

Yn dit proefskrift wurdt dy semantyk ta útgongspunt nommen. Ik bestudearje en ûntwikkelje útwreidings fan dy semantyk nei twa kanten: dy wurdt 'multyaktor' en 'dynamysk' makke.

Yn it tredde haadstik bestudearje ik in útwreiding fan epistemyske logika nei de kant fan: der is net ien operator $\Box$, mar der binne ferskate operatoaren fan de foarm $\Box_a$ yn de taal oanwêzich, foar eltse aktor a ien. In sin fan de foarm $\Box_a\phi$ betsjut dan: 'aktor a leaut dat ϕ wier is'. De semantyk is deselde as dy fan $\Box$, útsein dan dat der no mei eltse aktor a in eigen tagonklikheidsrelaasje yn it model korrespondearret. Dy semantyk makket it mooglik in tal nije epistemyske operatoaren te definiearjen, dy't, sawol fan in filosofysk as fan in logysk eachweid út besjoen, nijsgjirrich binne. Operatoaren foar mienskiplik witten en distribuearre witten binne al wol bekend út 'e literatuer. De definysjes fan dy operatoaren roppe fragen op oer de status fan de modellen; dy fragen komme yn besprek. Dy beide logika's hawwe ek nijsgjirrige bewizen oangeande folsleinens. Wat nij is yn dit proefskrift, is it konsept fan 'kombinearre witten'; dat liket faaks wol hiel bot op distribuearre witten, mar falt der dochs net mei gear. Ik besprek ek hiel yn 't koart hoe't jo de betsjutting fan dy operatoaren ôfhinklik meitsje kinne fan it ûnderwerp dêr't se oer geane.

De beide dêroanfolgjende haadstikken, 4 en 5, geane oer de útwreiding fan epistemyske logika mei epistemyske aksjes. Ik foegje oan de taal 'programma's' ta dy't feroarings yn de ynformaasje fan de aktoaren útdrukke. Sa is der bygelyks in programma dat útdrukt dat a en b beiden leare dat c leart dat ϕ wier is. Yn it fjirde haadstik definiearje ik in semantyk foar dy útwreiding fan de taal en formulearje ik in folslein aksiomasysteem foar de logika dy't dêrút fuortkomt. Dy logika hat in tal nijsgjirrige eigenskippen; ien dêrfan is dat as jo leare dat in sin wier is, soks net altiten ynhâldt dat jo yn in tastân telâne komme fan ek te leauwen dat dy sin wier is. Yn it fyfte haadstik besprek ik it wurk fan in tal auteurs dy't oer it ûnderwerp 'feroaring fan ynformaasje' skreaun hawwe, en ik ferlykje harren wurk mei de semantyk dy't ik yn haadstik 4 definiearre haw.

Yn 'e beide lêste haadstikken besykje ik sjen te litten hoe't de semantyk dy't ik yn dit proefskrift útienset haw, fan nut wêze kin foar in formele analyze fan dialogen. Yn it seisde haadstik besprek ik in tal aspekten fan it konsept 'common ground' sa't dat brûkt wurdt yn 'e literatuer oer dialogen. Om krekter te wêzen: dat haadstik giet oer de relaasjes tusken feroarings yn de common ground en yn de ynformaasje fan elts dy't deryn behelle is.

Yn it lêste haadstik besprek ik twa puzzels dy't te krijen hawwe mei feroaring fan ynformaasje: de puzzel fan de smoarche bern en de paradoks fan it ûnferwachte eksamen. Allebeide kinne jo krekt en ienfâldich beskriuwe yn de taal dy't yn dit proefskrift definiearre is. De observaasje dat it leard hawwen fan in sin net altiten ynhâldt dat jo dy sin ek foar wier oannimme, spilet yn dy útlis in wichtige rol. It haadstik wurdt ôfsletten mei in beskriuwing fan in ienfâldich dialoochspultsje.

De earste beide haadstikken fan it proefskrift haw ik noch net neamd. Dy haadstikken hawwe in wat technysker aard en hawwe as doel om de teoretyske eftergrûn te jaan foar de formele metoaden dy't ik yn dit proefskrift brûk. Yn stee fan de standertteory fan samlings haw ik keazen foar it brûken fan in samlingeteory dy't net wol-fundearre is, om dêrneffens epistemyske tastannen modellearje te kinnen. Mei't dy teory net sa bekend is, like it my goed ta dit proefskrift mei in algemiene yntroduksje yn dy teory te begjinnen. De wichtichste reden om dy net-standert metoade te brûken is dat dy de mooglikheid iepenet operaasjes dy't modellen feroarje te definiearjen, op in wize dy't direkter en eleganter is.

Modellen foar epistemyske logika dy't net wol-fundearre binne, ferskille fan de mear wenstige Kripkemodellen. It giet der dan om dat it by de lêstneamde mooglik is om beskate ûnderskiedings te meitsjen dy't jo net by de earste meitsje kinne. It komt derop del dat by it brûken fan net wol-fundearre modellen de ferskillen tusken bisimulêre modellen weifalle. It twadde haadstik omfettet in samling resultaten oer bisimulaasje en de relaasje dêrfan mei epistemyske logika. Meiinoar meitsje hja it oannimlik dat it weifallen fan it ûnderskied tusken bisimulêre modellen sûnder swierrichheden oan foarby gien wurde kin.

Titles in the ILLC Dissertation Series:

ILLC DS-1993-01: **Paul Dekker**
Transsentential Meditations; Ups and downs in dynamic semantics

ILLC DS-1993-02: **Harry Buhrman**
Resource Bounded Reductions

ILLC DS-1993-03: **Rineke Verbrugge**
Efficient Metamathematics

ILLC DS-1993-04: **Maarten de Rijke**
Extending Modal Logic

ILLC DS-1993-05: **Herman Hendriks**
Studied Flexibility

ILLC DS-1993-06: **John Tromp**
Aspects of Algorithms and Complexity

ILLC DS-1994-01: **Harold Schellinx**
The Noble Art of Linear Decorating

ILLC DS-1994-02: **Jan Willem Cornelis Koorn**
Generating Uniform User-Interfaces for Interactive Programming Environments

ILLC DS-1994-03: **Nicoline Johanna Drost**
Process Theory and Equation Solving

ILLC DS-1994-04: **Jan Jaspars**
Calculi for Constructive Communication, a Study of the Dynamics of Partial States

ILLC DS-1994-05: **Arie van Deursen**
Executable Language Definitions, Case Studies and Origin Tracking Techniques

ILLC DS-1994-06: **Domenico Zambella**
Chapters on Bounded Arithmetic & on Provability Logic

ILLC DS-1994-07: **V. Yu. Shavrukov**
Adventures in Diagonalizable Algebras

ILLC DS-1994-08: **Makoto Kanazawa**
Learnable Classes of Categorial Grammars

ILLC DS-1994-09: **Wan Fokkink**
Clocks, Trees and Stars in Process Theory

ILLC DS-1994-10: **Zhisheng Huang**
Logics for Agents with Bounded Rationality

ILLC DS-1995-01: **Jacob Brunekreef**
On Modular Algebraic Protocol Specification

ILLC DS-1995-02: **Andreja Prijatelj**
Investigating Bounded Contraction

ILLC DS-1995-03: **Maarten Marx**
Algebraic Relativization and Arrow Logic

ILLC DS-1995-04: **Dejuan Wang**
Study on the Formal Semantics of Pictures

ILLC DS-1995-05: **Frank Tip**
Generation of Program Analysis Tools

ILLC DS-1995-06: **Jos van Wamel**
Verification Techniques for Elementary Data Types and Retransmission Protocols

ILLC DS-1995-07: **Sandro Etalle**
Transformation and Analysis of (Constraint) Logic Programs

ILLC DS-1995-08: **Natasha Kurtonina**
Frames and Labels. A Modal Analysis of Categorial Inference

ILLC DS-1995-09: **G.J. Veltink**
Tools for PSF

ILLC DS-1995-10: **Giovanna Cepparello**
Studies in Dynamic Logic

ILLC DS-1995-11: **W.P.M. Meyer Viol**
Instantial Logic. An Investigation into Reasoning with Instances

ILLC DS-1995-12: **Szabolcs Mikulás**
Taming Logics

ILLC DS-1995-13: **Marianne Kalsbeek**
Meta-Logics for Logic Programming

ILLC DS-1995-14: **Rens Bod**
Enriching Linguistics with Statistics: Performance Models of Natural Language

ILLC DS-1995-15: **Marten Trautwein**
Computational Pitfalls in Tractable Grammatical Formalisms

ILLC DS-1995-16: **Sophie Fischer**
The Solution Sets of Local Search Problems

ILLC DS-1995-17: **Michiel Leezenberg**
Contexts of Metaphor

ILLC DS-1995-18: **Willem Groeneveld**
Logical Investigations into Dynamic Semantics

ILLC DS-1995-19: **Erik Aarts**
Investigations in Logic, Language and Computation

ILLC DS-1995-20: **Natasha Alechina**
Modal Quantifiers

ILLC DS-1996-01: **Lex Hendriks**
Computations in Propositional Logic

ILLC DS-1996-02: **Angelo Montanari**
Metric and Layered Temporal Logic for Time Granularity

ILLC DS-1996-03: **Martin H. van den Berg**
Some Aspects of the Internal Structure of Discourse: the Dynamics of Nominal Anaphora

ILLC DS-1996-04: **Jeroen Bruggeman**
Formalizing Organizational Ecology

ILLC DS-1997-01: **Ronald Cramer**
Modular Design of Secure yet Practical Cryptographic Protocols

ILLC DS-1997-02: **Nataša Rakić**
Common Sense Time and Special Relativity

ILLC DS-1997-03: **Arthur Nieuwendijk**
On Logic. Inquiries into the Justification of Deduction

ILLC DS-1997-04: **Atocha Aliseda-LLera**
Seeking Explanations: Abduction in Logic, Philosophy of Science and Artificial Intelligence

ILLC DS-1997-05: **Harry Stein**
The Fiber and the Fabric: An Inquiry into Wittgenstein's Views on Rule-Following and Linguistic Normativity

ILLC DS-1997-06: **Leonie Bosveld - de Smet**
On Mass and Plural Quantification. The Case of French 'des'/'du'-NP's.

ILLC DS-1998-01: **Sebastiaan A. Terwijn**
Computability and Measure

ILLC DS-1998-02: **Sjoerd D. Zwart**
Approach to the Truth: Verisimilitude and Truthlikeness

ILLC DS-1998-03: **Peter Grunwald**
The Minimum Description Length Principle and Reasoning under Uncertainty

ILLC DS-1998-04: **Giovanna d'Agostino**
Modal Logic and Non-Well-Founded Set Theory: Translation, Bisimulation, Interpolation

ILLC DS-1998-05: **Mehdi Dastani**
Languages of Perception