Playing with Information

ILLC Dissertation Series DS-2010-02

INSTITUTE FOR LOGIC, LANGUAGE AND COMPUTATION

For further information about ILLC-publications, please contact

Institute for Logic, Language and Computation
Universiteit van Amsterdam
Science Park 904
1098 XH Amsterdam
phone: +31-20-525 6051
fax: +31-20-525 5206
e-mail: illc@science.uva.nl
homepage: http://www.illc.uva.nl/

Playing with Information

ACADEMISCH PROEFSCHRIFT

ter verkrijging van de graad van doctor aan de
Universiteit van Amsterdam
op gezag van de Rector Magnificus
prof.dr. D. C. van den Boom
ten overstaan van een door het college voor
promoties ingestelde commissie, in het openbaar
te verdedigen in de Aula der Universiteit
op dinsdag 9 maart 2010, te 14.00 uur

door

Jonathan Alexander Zvesper

geboren te Norwich, Verenigd Koninkrijk
van Groot-Brittannië en Noord-Ierland.

Promotiecommissie

Promotor: prof.dr. K.R. Apt
Promotor: prof.dr. J.F.A.K. van Benthem

Overige leden:
dr. Alexandru Baltag
prof.dr. Jan van Eijck
prof.dr. Peter van Emde Boas
prof.dr. Dov Samet
prof.dr. Frank Veltman
dr. Yde Venema
prof.dr. Rineke Verbrugge

Faculteit der Natuurwetenschappen, Wiskunde en Informatica
Universiteit van Amsterdam
Science Park 904
1098 XH Amsterdam

Research supported by the 'research training host fellowship' GLoRiClass of the European Commision, MEST-CT-2005-020841.

Cover design by the author based on a photograph by Alessandra Lombardo.

Printed and bound by Copytech (UK).

ISBN: 978–90–5776–206–2

Contents

Acknowledgements

"*[R]ationality of thought imposes a limit on a person's concept of his relation to the cosmos*" – John F. Nash Jr., [1995]

Krzysztof Apt and Johan van Benthem have had the unenviable task of advising a regularly un-punctual and sometimes humourless sloppy thinker for three years. It is difficult to imagine how they managed to do this, but to do so while still being willing to talk to me, and indeed to offer outstanding advice comments and criticism, requires further leaps. I thank them for generously sharing their ideas and synergistic perspectives on the topics covered here and beyond; for their immense patience; and for so many inspiring meetings.

Thanks also to all my committee members; everybody who sent comments sent some very useful comments, to which I hope to have done justice in the subsequent revisions. Additional thanks to Alexandru Baltag for all of our discussions, and for his support and encouragement.

I'd like to thank the instigator, participants, invited speakers and co-organisers of the various Palmyr workshops.[1]

Many people made the ILLC such a great place to work and play, I owe you all something, and to many of you much. I am grateful to Krister Segerberg and Eric Pacuit, as well as my supervisors, for teaching me some of the, by their standards doubtless fairly elementary, mathematical skills required to write this thesis. For the especially good times, academic and otherwise, I offer my acknowledgements to Olivia Ladinig, Cédric Dégremont, Gaëlle Fontaine, Raul LeaL, Amélie Gheerbrant, Andi Witzel, Witzel Yun Qi, Simon Pauw, Umberto Grandi, Sujata Ghosh, Jacob Vosmaer, Nina Gierasimczuk, Jakub Szymanik, Henrik Nordmark, Daisuke Ikegami, René Goedman, Reut Tsarfaty and Joel and Sara Uckelman. Wouter (Koolen-Wijkstra): bedankt voor de op-het-laatste-moment samenvatting!

[1] `http://www.illc.uva.nl/PALMYR`

Katie K tirelessly read my woes over the years, and deserves some kind of honorific, possibly made of jam. Rafe: I'm sorry that my academic work took precedence over other projects. Danda provided the photograph on which the front cover's image is based, and wonderful energy. everyone at Camp Busted, and Carla and Dannette, helped keep me relatively sane during the penultimate writing stages. Nadia didn't-dance one Sunday morning. Louis, Thomas, Emma and Brad are the best imaginable sibling-like family members. And I am grateful to my parents for each selflessly doing so much for me.

Oxford, J. A. Z.
January 2010.

Introduction

"We don't have much left to do, we British, except to play our games" – Methwold [Rushdie, 1981]

Allow me to retain for a few more sentences a chatty tone, and the first person. Taking a step back from this Thesis, there is so much more I would have liked it to be, so let us say, while also retaining a down-to-earth sense of proportion, that it is aimed towards a theory of interactive reasoning. Game theory is for the most part the object of its study, and with good reason: game theory is *all about* interaction. But game theory needn't have a monopoly here. Computer scientists and philosophers are increasingly studying aspects of interaction too, with some formality [Dégremont and Zvesper, 2010], and something this thesis does is bring some aspects of their common ground – *logic* – to the fore.[2] We're still in chatty mode (just wait until you get to Chapter 1 if you want to see what 'dry' means), so we're allowed to admit some weaknesses. Sometimes we will get 'bogged down' in details of the particular disciplines we're writing about. We prove certain Propositions and Theorems that might leave people cold, and our discussion could sometimes seem wide of the mark. People from game theory will say, 'reduction axiom what now?', and people from logic might wonder why we care about 'players' (as in real players, not ∀belard and ∃loise). Still, bear in mind that we're contributing to something interdisciplinary here, all with the hope that just *one* truly interesting idea about interactive reasoning will eventually be squeezed out of the confusion.

In this Introduction we will first present, with minimal technical baggage or machinery, the sort of things that are discussed in the following Chapters. So we talk a bit about game theory, and then about interactive epistemology (slipping in a little bit of logic). Those few pages are *not* intended as a guide for the details of what is in the

[2] A shift of focus to interaction can be said to have been brewing for several decades, as a concerted effort to change the fact that "[t]raditional philosophy of language, like much traditional philosophy, leaves out other people and the world" [Putnam, 1975, p. 193].

rest of this Thesis, but rather to whet the (relatively) lay reader's appetite. We will also explain the so-called 'deductive' interpretation of game theory that we will principally have in mind throughout this Thesis, and give some reasons why we prefer to take 'belief' rather than 'knowledge' as the object of (interactive) epistemology. Then, after we've mentioned a few things we find interesting from the fields of game theory and interactive epistemology, we do endeavour to explain some of that will occur in the Chapters, one by one.

Game theory

Game theory is the mathematical study of the interactions of largely idealised decision-makers. Mathematical in the following sense: it abstracts from much of the detail of those interactions qua events taking place in the real world (which "might better be called the complex world" [Aumann, 1985]). The advantage to such an abstraction being that game theorists can present formal models, about which they can prove theorems. These theorems, in turn, are supposed to tell us something about the complex world we abstracted away from.

To quote from the opening passage of a popular textbook on game theory [Osborne and Rubinstein, 1994],

> "*The basic assumptions that underlie [game] theory are that decision-makers pursue well-defined exogenous objectives (they are* rational*) and take into account their knowledge or expectations of* other *decision-makers' behaviour (they* reason strategically*).*"

The Thesis you are reading uses formal tools drawn principally from work in philosophical logic to explore both of these notions, *rationality* and *strategic reasoning*, focusing especially on the role of information and belief: on what are called 'epistemic' or sometimes 'doxastic' aspects.

Rationality here is taken to mean '*instrumental*' rationality, i.e. rationality means just that players pursue their objectives, that can be taken to represent their 'best interests', as perceived by themselves. Players therefore function as optimisers of *what they take to be* their own best interests. There is a doxastic component therefore in the definition of this most fundamental notion in game theory:

> "*A person's behaviour is* ***rational*** *if it is in* ***his*** *best interests, given* ***his*** *information.*"

That is the standard definition of (instrumental) rationality; the particular quotation is from [Aumann, 2006].

Perhaps the most widely known example of a game (in the sense of game theory) is the so-called "prisoner's dilemma". One formulation, from [Osborne and Rubinstein, 1994, p. 16], of the prisoner's dilemma is the following:

> *Two suspects in a crime are put into separate cells. If they both confess, each will be sentenced to two years in prison. If only one of them confesses, he will be freed and used as a witness against the other, who will receive a sentence of three years. If neither confesses, they will both be convicted of a minor offence and spend one year in prison.*

We will not introduce formally the definitions of games, including of players' preferences (or 'utilities'), until Chapter 1, but for present purposes allow us to represent the situation using the matrix in Figure 1. This matrix represents what we will call a

	Y	N
Y	$1, 1$	$3, 0$
N	$0, 3$	$2, 2$

Figure 1: A matrix representing the prisoner's dilemma game

'strategic game'; the important point about it is that it is intended to capture the essential parts of the given description of the situation the players (the prisoners) find themselves in. One player (the 'row player') must choose between the top and bottom rows, and the other (the 'column player') must choose between the left and right columns. The numbers in the resulting entry in the matrix (e.g. $0, 3$) then represent the 'utility' obtained by the row and column players respectively. The choice Y (the top row for the row player; left column for the column player) represents confessing to the crime; the choice N represents not confessing.[3] The utilities written in the boxes are made on the assumption that the only concern the players have is how much time they will spend in prison: 0 means spending three years in prison, 1 means spending two years, etc.

Players prefer higher utilities, which means that in this case mainstream game theory makes a unique prediction: both players will play D, i.e. both players will confess. However, there are two ways to think about that prediction, which rely respectively on the '*deductive*' and the '*steady-state*' interpretations in game theory (cf. [Osborne and Rubinstein, 1994, Section 1.5]). The ***deductive interpretation*** will suit this scenario better; it says that a game matrix really represents a 'one-shot' interaction, in which players use only reasoning about the game, with no exogenous information. According to the deductive interpretation, players can perform the following kind of reasoning. The row player can say, 'if my fellow prisoner plays Y then I would be better off playing Y; and if my fellow prisoner plays N then I would be better of playing Y; so I should play Y'. Thus N is, to use game-theoretic jargon that we introduce in Chapter 1, 'strictly dominated' by Y.

The ***steady-state interpretation*** of game theory is very different, and is the interpretation that supports the notion of 'Nash equilibrium'. A Nash equilibrium is a 'profile'

[3] These 'confessions' need not be sincere, as nothing in the scenario says whether or not the suspects are guilty.

of strategies, i.e. one strategy for each player, such that *given* that those strategies are played, no player has an incentive to deviate from the profile. To put it otherwise, a Nash equilibrium is a best response to itself. In the prisoner's dilemma, the unique Nash equilibrium is again where both players play Y. – In any other entry in the matrix (i.e. in any other profile) at least one player has an 'incentive' to deviate from that profile. However, in a great many other games, the steady-state interpretation yields very different answers from the deductive interpretation.

For example, in so-called 'pure coordination games' like that in Figure 2, the deductive interpretation cannot make any prediction, since L and R are symmetric for both players. However, only (L, L) and (R, R) are Nash equilibria. From the point of

	L	R
L	$1, 1$	$0, 0$
R	$0, 0$	$1, 1$

Figure 2: A pure coordination game

view of the deductive interpretation then, the steady-state interpretation makes some kinds of additional assumptions, that arguably should be integrated into the description of the game. The steady-state interpretation is sometimes taken to assume tacitly some notion of *repetition* of the scenario being represented.[4] However, repetition itself substantively changes the game[5]. So perhaps communications or signals of some kind, for example those underlying Aumann's [1974] notion of correlated equilibrium, might be the best way to understand the steady-state interpretation of game theory.

Our interest almost throughout this work will be focused on the conceptually clearer *deductive* interpretation of game theory. So we in general have in mind a 'one-shot' kind of interaction, in which any repetition or communication should be modelled explicitly as part of the game. (Furthermore, let us remark that we will not have anything to say about *cooperative* game theory.) Within the deductive interpretation, we will look at 'interactive epistemology', that is reasoning about *beliefs*, including about beliefs concerning beliefs.

Formal interactive epistemology

Alongside mathematical structures that represent the games themselves, we will consider mathematical structures that are intended to formalise the notion of *information*

[4]Sometimes the assumption is made more more explicit: "as a given setting gets more and more common and familiar, it makes [the players] act more and more rationally in that setting" [Aumann, 1985].

[5]The observation that an *arbitrarily repeated* prisoner's dilemma yields a different outcome was what won Aumann the Nobel prize in economics [Aumann, 2006].

or *belief*, and so to get a handle on this talk of strategic reasoning, and indeed of rationality. These structures, or 'models', represent "knowledge or expectations" of the players.

There is more to rationality and to strategic reasoning than simply having expectations about other players' behaviour. Those expectations themselves, e.g. player i's beliefs about what player j will do, are often derivable from, and must at least be consistent with, some more fundamental beliefs of player i. For instance player i might herself make the "basic assumption" mentioned in the quotation from Osborne and Rubinstein, so that i would in particular believe (since it follows from her "assumption") that player j is rational and reasons strategically, perhaps based on a similar assumption.

We prefer to use the term 'belief' rather than 'knowledge' for a number of reasons. Firstly, we sometimes want to allow for the information the players have to be incorrect. (Or if information has by definition to be correct, then we are allowing for what the players *think* is their information to be incorrect.) Furthermore, we prefer the conceptual position which holds that given that the game itself somehow represents 'real' possibilities, the players do not *know* of any of the possibilities that it will not occur: if you *know* that your opponent will not play a certain move, then arguably that move should not be included.[6] Using the terminology of Brandenburger [2007], we take a *belief-based approach* to game theory, that he outlines as follows:

> "*[O]nly observables are knowable. Unobservables are subject to belief, not knowledge. In particular, other players' strategies are unobservables, and only moves are observables.*" (op.cit., p. 489)

However, this choice of ours need not be taken to reflect any deep philosophical or epistemological point, and much of what we say about belief will also hold for knowledge. One almost indisputable property of knowledge, that clearly distinguishes it from belief, is that it is a 'factive mental state operator' [Williamson, 2000]: that if one knows something, then it is true. Plato's definition of knowledge, as justified true belief, has been shown to be wanting by Gettier's famous counterexamples, but it is certainly not controversial to maintain that it is a necessary condition, for a belief to be knowledge, that it be true.

If we were to insist that all beliefs modelled were true, perhaps we could call them 'knowledge'. Indeed, if the reader particularly likes the term 'knowledge', and is unpersuaded by the above-cited view of Brandenburger, then she can substitute it for 'true belief' wherever she likes. Most of the results that we establish for belief hold for always-true belief and so, for such a reader, for knowledge. (Indeed everything in Chapters 1 and 2 holds reading 'knowledge' in the place of 'true belief'; Chapters 3 and 4 make more fundamental use of the belief-based approach.)

A formal logical approach to studying the notions of knowledge and beliefs was instigated by Hintikka [1962], using so-called 'modal logic'. And game theorists, most

[6] To fully motivate this line of argument we would have to say that if there is 'common knowledge' that s will not be played then there is no reason to include s in the description of the game.

notably Aumann [1976], have independently developed formal models for knowledge and beliefs along the same lines.

Some crucial limitations to Hintikka's work have since been overcome. For example, Hintikka writes that "I do not know how to characterize the notion of occasion exhaustively" (op.cit., p. 7), whereas the *semantics* that were soon to be developed for modal logic (pre-empting the epistemic models introduced in the game-theoretical literature) furnish precisely such a characterisation. The logical approach, and the knowledge and belief models proposed in the game theory literature, were for a long time what we will call 'static'. That is, "[t]here cannot be any question of increasing one's factual knowledge"; what is more, the only assertions about beliefs or knowledge that are susceptible to the formal analysis proposed there are those "made on *one and the same occasion*" ([Hintikka, 1962]). Yet Sorensen [2009] is able to write that

> "*just as it is easier for an Eskimo to observe an arctic fox when it moves, we often get a better understanding of the knower dynamically, when he is in the process of gaining or losing knowledge.*"

Even if that quotation does describe the situation a little too colourfully (metaphorically speaking), still the *change* of beliefs and knowledge is an important phenomenon, and we will relate it to our study of games.

An important concept in interactive epistemology is that of 'common knowledge', which we can think of as a special case of 'common belief'. A fact is commonly believed (by a group) if everybody (in that group) believes it, they all believe that they all believe it, and so on. (Actually we will see in Chapter 1 that this 'and so on' hides some subtleties.) [1976] presents a formalisation of common knowledge. The concept had already been discussed in [Lewis, 1969] and indeed formalised in [Friedell, 1969] (under the name 'common opinion').[7]

We just saw a very small game, prisoner's dilemma, in which both players can, on the basis of their rationality alone, eliminate strategies and so arrive at a conclusion of what they will do. So in that game, any information that the players might have concerning the rationality of the other player is entirely irrelevant for them to decide how to play. Now consider the slightly larger game in Figure 3. Here a similar piece of

	L	C	R
U	$2,2$	$0,1$	$2,0$
M	$1,3$	$2,2$	$2,1$
D	$0,0$	$1,3$	$3,1$

Figure 3: A game where higher-order beliefs about rationality are important

reasoning as in the case of prisoner's dilemma (Figure 1) means that the column player,

[7] In [Aumann, 1976] the author was apparently unaware of these earlier works, and can be credited with bringing the importance of the concept to the attention of game-theorists. Dov Samet drew our attention to [Friedell, 1969]; maybe it will soon be common belief who first formalised common belief.

b, will not play R if he is rational. For *no matter what* his opponent a does (either U, M or D), he would be better off playing C than R. That is, C *strictly dominates* R. In the kind of formal notation that we will use in Chapter 2, this fact, that b's rationality entails not playing R, might be written as:

$$\mathbf{r}_b \rightarrow \neg R. \tag{1}$$

The formula (1) is understood as being true *everywhere in any model* of the game, since it does not depend on any factors exogenous to the game.

What about player a, are there any similarly '*stupid*' moves for her? Not really: that is, none of U, M or D are strictly dominated. However, suppose that a has the information that b is rational. That is, in some formal notation:

$$\Box_a \mathbf{r}_b. \tag{2}$$

The formula (2) is not necessarily true everywhere in every model, since it is conceivable that player a might believe that player b is not rational. But (without going into detail of the definitions of different kinds of models, which are to be found in Chapters 1, 2 and 3), it can be true *somewhere* in a model, let's say at some 'state'. Suppose furthermore that a is *able to draw inferences* so that when some implication $A \rightarrow B$ is true everywhere in the model, and she (at some state) believes A, then she (at that same state) believes B (technically: if her belief modality is *monotonic*). Then clearly, at any state where (2) holds, we will have

$$\Box_a \neg R \tag{3}$$

That is: a has the *information* that b will not play R. But in this case, a's rationality means that she will not play D, since no matter what b plays that is compatible with a's information (i.e. L or C), a would be better of playing M.

So, writing $\wedge$ for 'and', we have

$$(\mathbf{r}_a \wedge \Box_a \mathbf{r}_b) \rightarrow \neg D \tag{4}$$

But this reasoning can go on, in the sense that if b believes that a is rational, *and* that a believes b is rational, we find that $\Box_b \neg D$, i.e. b now would have the information that a will play U or M. In which case b's rationality would mean not playing C.

Still we are not finished: it actually turns out that if players all are rational and commonly believe in each other's rationality, then in this game they can only play outcomes that survive the *iterated* elimination of strictly dominated strategies. In this particular game that means playing according to (U, L).

This sort of result, that illustrates the connection between the deductive approach to game theory and its epistemic analysis, has been established in [Bernheim, 1984; Pearce, 1984; Tan and Werlang, 1988].

Contributions we will make

That brief sample of ideas from game theory and from interactive epistemology can really only serve to whet the appetite: many more aspects to both topics are introduced throughout the various parts of this Thesis. We now try to summarise what we will add, in each of the Chapters that follow: what contribution each Chapter makes.

Chapter 1 In the first Chapter we will generalise some of the results from the literature we just mentioned, relating ***mutual belief*** in ***rationality*** with the ***iterated elimination*** of non-optimal strategies. Section 1.1 introduces all of the technical definitions related to ***strategic games*** and ***optimality operators***. Section 1.2 gives a full heuristic treatment of the rest of the Chapter, avoiding as much as possible technical details. We find that to be necessary because there are a number of subtleties involved. Still let us attempt to summarise here what we will do in the technical part of the Chapter. First of all, we generalise the result mentioned above about common belief of rationality to the ***infinite*** case (Theorems 1.1 and 1.2). Then we consider, as did Tan and Werlang [1988] for the finite case, ***arbitrary stages*** along the way to full common belief. This involves employing a distinction between two different forms of common belief, and borrowing from the literature non-standard ***'neighbourhood' models*** of beliefs in order to distinguish for example between mutual belief to depth ω_0 and to depth $\omega_0 + 1$, where ω_0 is the first infinite ordinal. We use the fact that we can make this kind of distinction in neighbourhood models to show that there is a model where for every stage, including the transfinite stages, of iterated elimination of non-optimal strategies, there is some information that 'rationalises' it. That is Theorem 1.5, where the model we provide is actually a ***topological*** neighbourhood model, meaning that the only difference between it and a standard model is that players might ***fail to put together*** *large* amounts of information. That is, they might have many pieces of information $\varphi_1, \varphi_2, \ldots$, and thereby also have all *finite* implications of this information, while still failing to draw all the conclusions that might be possible when considering *all* the ϕ_n's.

More generally, neighbourhood models allow for the case where a player does not put her information together, even finitarily. So for example she might believe that φ and believe that $\varphi \rightarrow \psi$ (that φ entails ψ) and still not believe that ψ. Equivalently: neighbourhood models allow for a situation where a player believes φ and believes ψ without believing their conjunction $\varphi \wedge \psi$. Thus they are even more 'permissive' than topological models, which only allow that players fail to put together infinite amounts of information. Neighbourhood models therefore provide some way to model imperfection of reasoning, where reasoning might be constrained by the nature of the player who for example does not have time to put together her information. In any case, we show that under certain rather weak conditions about ***introspectivity of beliefs***, even this kind of neighbourhood models are enough to prove the kind of result we obtained already for the relational model case. The two different conditions we consider yield two Theorems: 1.3 and 1.4.

Chapter 2 In the next Chapter we introduce formal logical languages, like those used by Hintikka, for reasoning about beliefs. Taking as a starting point arguments from Aumann [1999], we look at a number of reasons for using formal languages in epistemic analysis: for making a distinction between ***syntax*** and ***semantics***. One of the arguments that we give in favour of using modal languages in game-theoretic analyses, is that these languages are appropriately ***local***. We catalogue many choices that can be made at the level of the language, usually sticking within the realm of modal languages, though some of them are just notational variants of for example first-order logic. We address the question of ***definability*** of key notions from game theory, like rationality and common belief. We also spend considerable space on a foundational question concerning the existence of a suitably 'large' belief model. That is, we study the property of '***assumption-completeness***' introduced in [Brandenburger and Keisler, 2006]. This leads us to introduce the '***type-space models***' used in that work, and to show how they are related to the more standard state-space models. A two-player type-space model is assumption-complete for a language if for every sentence of that language that defines a set B of b's types, there is an a type where a's information is precisely B. (Assumption-completeness is related to the 'comprehension schema' in set theory.) We examine what assumption-completeness means in state-space models. The principal technical contribution of the Chapter is to prove (Theorem 2.4) that for *infinitary modal languages* there are assumption-complete models.

One of the arguments that we give in favour of using a logical language is that this facilitates reasoning about events across different models, which is very useful in introducing ***dynamics***, in any field but in this case into the study of games. The next two Chapters introduce dynamics into the picture.

Chapter 3 In the first of them, we discuss dynamic epistemic logic, and extend some results to cover the case of neighbourhood models. That Chapter then returns to strategic games, and explicitly formalises some interactive reasoning process that are compatible with the deductive interpretation of game theory. This is one role played by belief dynamics: as a metaphor for the reasoning or computation that is involved in arriving at conclusions about games. We can think of the game as specifying an initial epistemic or informational state, further epistemic states being induced by reasoning from premises saying that the players are rational, reasoning of which we also give a logical account. We interpret this as a ***private but common*** reasoning process. This attempt to tell a coherent story about the deductive process leads us to look not only at the 'hard' information case but also at 'soft' information, i.e. to consider ***revisable*** beliefs. We introduce the notion of a '***rational equilibrium of beliefs***', by which we mean a configuration of beliefs that is stable to further deduction, and we argue that in general using soft information (and so revisable beliefs) is the only way to arrive at a rational equilibrium of beliefs, at least in the case of some ***non-monotonic*** optimality properties.

Chapter 4 In the last Chapter we turn our attention to applying logical analyses to epistemic aspects of ***extensive-form games***. In an extensive form game players do not, as in the case of strategic games, make their choices entirely independently of the choices made by the other players. That is, an extensive-form game represents a decision process that is extended in time, with players making choices one after the other. The crucial difference in terms of our concerns about beliefs is that the beliefs of the players can change *as the game is played*. The main contribution we make in that Chapter is to offer an analysis of ***backward induction*** in terms of beliefs. In backward induction, players reason about what would happen hypothetically, and in a large class of games (including so-called 'generic' games in which no player is indifferent between two different outcomes), this purely deductive reasoning will yield a unique prediction for the game. However, it has been a thorny question exactly what configuration of beliefs or knowledge is required in order to guarantee that players will play according to the backward induction prediction. We offer (Theorem 4.1) such conditions, phrased in terms of dynamics of revisable beliefs, and making crucial use of a notion of ***stability*** of belief, and a forward-looking ***'dynamic' rationality***.

A similar notion of 'rational equilibrium of beliefs' arises in this context, and we use this notion to reason about a simplified version of so-called trembling-hand perfect equilibrium, that we call ***even-handed***, that is a refinement of the usual notion. We suggest that ***belief revision policies***, in concert with lexicographic rationality, are a useful way to think about various solution concepts. Finally, we close the last Chapter by pointing to some limitations of our existing analysis of extensive games in terms of dynamic epistemic logic, specifically that it does not yet give a coherent account of ***strategic communication***.

Origin of the material

This work integrates and builds upon some of my major collaborations over the last three years, when it has been a privilege, as well as very enjoyable, to work with co-authors whom I would like to thank deeply. All errors of presentation and content naturally remain my responsibility.

- Most of the ideas from Chapter 1, and Theorems 1.1 and 1.2, are drawn from [Apt and Zvesper, 2007]. Theorems 1.3, 1.4 and 1.5 build on that collaboration but are original contributions.

- Some parts of Section 2.3 are drawn from [Zvesper and Pacuit, 2010], including Theorem 2.4, which is a generalisation, to the infinitary case, of (op.cit., Theorem 2.6).

- Much of Section 4.2, including Theorem 4.1, is drawn from [Baltag *et al.*, 2009]; furthermore, some of the ideas sketched in Section 4.3 are based on work in progress with the authors of that paper.

Chapter 1

Believing Rationality in Arbitrary Games

"To infinity, and beyond!"
– Buzz Lightyear
[Lasseter, 1995]

This Chapter examines mutual belief of rationality in one-shot interaction situations. Like all but parts of the last Chapter of this Thesis, this Chapter is concerned with a purely deductive interpretation, rather than with any element of steady-state interpretation, of game theory. So we consider what conclusions players can draw from a relatively minimal amount of information. That information will concern just the (instrumental) rationality of the other players, (where, recall, a player is instrumentally rational just if she acts in her best interest according to her information), and higher-order information about that information.

Thus we look at what it means for players to be rational, and to believe that the other players are rational, to believe that the other players believe that the other players are rational, etc. The most substantial contributions of this Chapter are to *generalise* some standard results from the game-theoretical literature, that connect the different levels of mutual belief in rationality with numbers of rounds of elimination of sub-optimal strategies. That generalisation has three parts to it:

1. As we explain in a moment, our theorems cover a broad class of optimality notions.

2. They also cover *infinite* games, where the results in the literature generally look at finite games.

3. Finally, we consider a larger class of models for beliefs, which means that we make very few assumptions about the ways players put their beliefs together. In terminology that we introduce later in the chapter:

(a) We allow for the case of 'relational' belief models in which players need not be 'positively or negatively introspective'.

(b) We also allow for the more general case of 'neighbourhood' belief models, in which players not only lack those introspection properties, but also do not necessarily 'put their beliefs together', i.e. believe all the things that follow from their beliefs.

If everybody believes some proposition E, then we say that there is ***mutual belief*** of E'; if everybody believes that everybody believes that E, we say there is ***second-level mutual belief*** of E, and so on. If there is mutual belief of E on all levels, this is called ***common belief*** of E.[1] As we will see, this definition can be made formal in two ways, depending on whether one includes only all *finite* levels of mutual belief, or *arbitrary* levels of belief, including levels for *transfinite ordinals*. That distinction is not usually made in the game-theoretical literature, and the models for beliefs commonly used there do not allow for the distinction to be made. Aumann [1976] was the first to formalise a notion of common knowledge (or as we might say: common true belief), and in his framework of 'partition structures' the distinction cannot be made, nor can it be made in the more general case of 'relational models'. However, it is possible to make this distinction in other, yet more general, 'neighbourhood' and 'topological' models for beliefs. We will exploit this distinction when we look at different levels of mutual belief of rationality in infinite games.

Rationality can be defined in many different ways, depending on what notion of 'optimality' is used by the players. In turn those different notions of optimality induce operators that reduce the game matrix by eliminating sub-optimal strategies. The first way in which our results are a generalisation of existing ones is that they are phrased not in terms of a specific optimality operator but always in more general terms.

So the results that we will prove all establish, roughly speaking, something of the form:

($\star$) Rationality plus α-level mutual belief in rationality is equivalent to all players avoiding strategies that are eliminated within $1 + \alpha$ rounds of elimination of non-optimal strategies.

The second generalisation is that we allow for the possibility that there might be an infinite number of objects of choice for any of the players, i.e. we give results for games with *arbitrary* strategy sets. Thus when we write α above, we mean it to refer to an arbitrary (possibly infinite) ordinal.[2] We will show why this entails, for one direction of the 'equivalence' established by our theorems, considering neighbourhood

[1] What we call 'mutual belief' is sometimes called 'general belief' in the literature. Note that, according to our terminology, mutual belief is in general *not* the same thing as common belief.

[2] Finite ordinals are just natural numbers $1, 2, \ldots$. Transfinite ordinals are studied in set theory [Devlin, 1993], and their arithmetic is not the same as that for finite ordinals, so that in particular for infinite ordinals $\alpha = 1 + \alpha \neq \alpha + 1$. Therefore (as we explain in Section 1.2 below) it is crucial that we write '$1 + \alpha$' in formulating the various theorems we prove.

and topological models. This in turn means that we will have to define what it means, in neighbourhood models, for a player to be *rational*, i.e. to extend the existing definition from relational models.

One half of the the final generalisation involves showing that in relational (or indeed topological) models, no further 'introspection' properties are required of players in order to obtain the result. The other half again involves using neighbourhood models. In those each player might fail to 'put together' her pieces of information, or indeed to 'draw conclusions' from her information: formally, her information neighbourhoods need not be *closed for intersection*, and need not be *monotonic*. Here we present two options in order to get our equivalence. The first is to introduce the new notion of 'co-mutual' belief, that we show is enough on neighbourhood models to get ($\star$) with 'co-mutual belief' replacing 'belief' even if players do not have *any* introspection properties. The second is to show that with just one minimal introspection property, we can get the result on neighbourhood models.

Background literature

The starting point in game theory for our own small contributions here are [Bernheim, 1984; Pearce, 1984; Tan and Werlang, 1988]. Those papers each show the connection between mutual belief of rationality and the elimination of non-optimal strategies. All of them consider only *finite* games, and each focuses on only one type of optimality. The more abstract approach of arbitrary monotonic operators, and the generalisation to infinite games, is studied in [Apt, 2007a].

On the side of interactive epistemology, there was some work on formal epistemology in the modal logic tradition, started by [Hintikka, 1962]. Aumann brought the attention of game theorists to the notion of *common knowledge*, by providing an elegant formulation of it and theorem about it [Aumann, 1976]. As we have said, it turns out that there are different ways to define common knowledge for infinitary cases; this fact was first established by [Barwise, 1988], and discussed further in [Heifetz, 1999; Benthem and Sarenac, 2004].

Barwise's 'situation semantics' framework was shown in [Lismont, 1994] to be equivalent to using 'neighbourhood models', developed in [Scott, 1970], and discussed in the textbook [Chellas, 1980]. A modern logical model-theoretic approach to neighbourhood models is presented in [Hansen *et al.*, 2009]. [Heifetz, 1996] also studied common belief on neighbourhood models.

Topological models for modal logic, that we also use below, originate in the work of McKinsey and Tarski [1944], and are studied from a contemporary logical perspective in [Benthem and Bezhanishvili, 2007]. They are used for epistemic logic in [Benthem and Sarenac, 2004], where again the distinction between two different varieties of common belief is drawn.

Finally let us remark that since players in neighbourhood models do not necessarily put their information together, using neighbourhood models to represent players' beliefs is a partial way to address the problem of 'logical omniscience', i.e. the prob-

lem that players believe all logical validities. We do not pursue that connection further here, and so do not entertain either of the two classical ways of addressing logical omniscience: the use of so-called 'impossible worlds' [Hintikka, 1975] or the distinction between implicit knowledge, which is logically omniscient, and explicit knowledge which is not [Fagin *et al.*, 1995].

Organisation of the Chapter

In Section 1.1, we spend some time going over standard definitions for game theory, including of strategic games.[3] In that Section we do not yet make any novel contributions. We present there the 'optimality operator' approach, and we show how the optimality operators can be instantiated by a number of concepts familiar from game theory, including avoiding strategies that are strictly dominated, and so on. Each concept can induce a number of different optimality operators depending on some details, including whether we consider pure or mixed strategies, and so on. (We also discuss mixed strategies and the connection, sometimes made in the literature, between them and beliefs.)

Then in Section 1.2 we give, avoiding as much as possible technical details, an explanation of the theorems that we will prove in Sections 1.3 and 1.4. In Section 1.3 we introduce formally the *relational* models of belief, and mention the introspection properties often attributed to players. The theorems in that Section relate *common belief* of rationality to the iterated elimination of non-optimal strategies. So in Section 1.3 we consider *full common belief* of rationality, which corresponds to finishing the process of iterated elimination of non-optimal strategies. But the elimination algorithm works in a stage-wise fashion, and we are interested in finding correlates on the epistemic side for each state in the process. In Section 1.4 we therefore look at intermediate (possibly transfinite) stages. As we explain in more detail in the heuristic treatment in Section 1.2 leads us to use neighbourhood models for belief, this means we have to use neighbourhood models for belief, and we prove the mentioned correspondence between α-level mutual belief in rationality with $1 + \alpha$ rounds of elimination of non-optimal strategies.

To re-iterate: Section 1.1 mainly repeats material that could be familiar to the reader well-versed in game theory, so such a reader might prefer to skip that Section except for looking briefly at the definition of optimality operator (Definition 1.2) and outcome ordinal (Definition 1.3).

1.1 Strategic games and optimality operators

As a preliminary to the material in this chapter, we will make formal our talk from the Introduction of strategic games and game reduction operations.

[3] The small games we looked at in the Introduction were all strategic games. Strategic games are also sometimes called "games in normal form", for example by von Neumann and Morgenstern [Neumann and Morgenstern, 1944] in their foundational work on game theory, to which the field owes its existence.

Recall that strategic games are intended to represent one-shot simultaneous-choice interactions. So there will be a set N of players, and each player $i \in N$ will have a set of 'choices' or 'strategies' denoted T_i. These will be unanalysed, primitive objects in the definition of strategic game. The set of ***strategy profiles*** or ***outcomes***, denoted T, is then just the Cartesian product $\prod_{i \in N} T_i$: an outcome specifies what strategy each player chose.

The other ingredient will be the preferences, or 'utilities', of the players in N over the outcomes. We will allow strategic games to be defined with ***ordinal preferences*** or with ***cardinal preferences***. With *ordinal* preferences, we state that players have a consistent 'preference order' over all possible outcomes of the game. This boils down to saying that given two possible outcomes, they can say which one, if either, they prefer, in such a way that we cannot catch them out as preferring a over b, b over c and c over a. That players have consistent ordinal preferences is of course a non-trivial statement, but it is a little less drastic than assuming that players have *cardinal* preferences, which says that the players assign a particular real number (element of the continuum) to every possible outcome of the game. Cardinal preferences over a set of outcomes T naturally induce ordinal preferences: if the value i assigns to a is greater than i assigns to b then i prefers a over b. But cardinal preferences are strictly more expressive than ordinal preferences: clearly different cardinal preferences can induce the same ordinal preferences. (For example where T is $\{a, b\}$, if i assigns 2 to a and 3 to b this is ordinally equivalent to i assigning 0 to a and 300 to b.)

Nonetheless, ordinal preferences will be sufficient for almost all of our purposes, and are conceptually a little less questionable than cardinal preferences. Throughout this thesis we will prefer to talk about games with ordinal preferences, though sometimes (for example in the present Chapter when we will talk about optimality operators from the literature which involve mixed strategies) we are forced to talk of games with cardinal preferences. Furthermore, sometimes when defining a game it is easier to write down cardinal preferences than ordinal preferences, but they can be thought as simply a shorthand notation for what is really an ordinal preference relation.

Definition 1.1. Fix some set of players N.

1. A ***strategic game with cardinal preferences*** for N is a tuple $(T_i, \pi_i)_{i \in N}$, where T_i is player i's set of 'choices', also called her 'strategies', and $\pi_i : T \to \mathbb{R}$ is her 'payoff function'.

2. a ***strategic game with ordinal preferences*** (sometimes in this Chapter and the next just called a ***game***), is a tuple $(T_i, \geq_i)_{i \in N}$, where each $\geq_i$ is a total order relation over T.[4]

3. $T = \prod_{i \in N} T_i$ is the set of ***strategy profiles*** or ***outcomes***.

[4]I.e. a total transitive antisymmetric relation. We write $>_i$ for the strict version of the relation ($s >_i t$ iff $s \geq_i t$ & $t \not\geq_i s$).

There are two natural ways to define *subgames*. Firstly, as tuples $(S_i)_{i\in N}$ with $S_i \subseteq T_i$ of subsets of the strategy sets in the original game; these we will call ***subgames*** (we do not include the preference information in the definition of a subgame, so a subgame only makes sense as a game in the context of the original game of which it is a subgame). The second way would be to define them as subsets $S \subseteq T$ of the strategy profiles in the original game; these we will call ***restrictions***. Any subgame $(S_i)_{i\in N}$ defines a restriction: $\prod_{i\in N} S_i$. Conversely, it is only 'rectangular' subsets $S \subseteq T$ (restrictions) that are definable in this way. For example if the original strategies were $(\{U, D\}, \{L, R\})$, then the restriction $S = \{(U, L), (D, R)\}$ clearly is not definable by any subgame. Although in this Chapter we will mainly be interested only in rectangular restrictions, still we often use restrictions just because they are sometimes notationally easier.

We need to introduce a few useful pieces of standard notation for manipulating strategies and restrictions. For any player i, we write T_{-i} to mean $\prod_{j\in N-\{i\}} T_j$. And given any $s_i \in T_i$ and $s_{-i} \in T_{-i}$, by (s_i, s_{-i}) we mean the relevant element in T. Similarly, given $S_i \subseteq T_i$ and $S_{-i} \subseteq T_{-i}$, the expression $S_i \times S_{-i}$ denotes the relevant subset of T. Given a restriction S, for any player i, we write S_i to mean the set of i's strategies occurring in some profile in S:

$$S_i = \{s_i \in T_i \mid \exists s_{-i} \in T_{-i} : (s_i, s_{-i}) \in S\},$$

and also extend this notation to S_{-i} in the analogous way. Sometimes we will refer interchangeably to a rectangular restriction and its corresponding subgame, so we could write for example (when the set N of players is irrelevant or clear from the context) $(T, \geq)$ to refer to the game $(T_i, \geq_i)_{i\in N}$.

Let us adapt a motivating example from [Morgenstern, 1928] (actually we entirely change the story, but the message is similar). Suppose Sherlock Holmes and his nemesis Moriarty are on a train from London which will stop only at Canterbury and Dover. The latter has a gun and hopes therefore to catch the former, and so wants to alight at the same stop as him. Holmes on the other hand, who has no way to defend himself save his cunning, wishes to evade capture, and so wants to alight at a different stop. Apart from that, Holmes would prefer not to stay on the train very long, because if he evades Moriarty he would like to return to London that evening. Moriarty on the other hand hopes to escape to France, so staying on this train to Dover is his preferred option. We can describe the game as follows:

$$\begin{array}{l} N = \{h, m\} \\ T_h = T_m = \{D, C\} \\ (C, D) >_h (D, C) >_h (C, C) \sim_h (D, D) \\ (D, D) >_m (C, C) >_m (C, D) >_m (D, C) \end{array}$$

Here the players are h (Holmes) and m (Moriarty), the strategies for either are D and C (Dover or Canterbury), and so the outcomes are e.g. (C, C) they both alight in Canterbury, or (C, D) Holmes alights at Canterbury, evading Moriarty who remains on

the train until Dover. The preferences are faithful to the story, so that for example for Moriarty the best option is for both Holmes and he to alight at Dover so that the evil mastermind can shoot the detective and hop aboard a ferry to the Continent.

We could also denote the situation as a game with cardinal preferences as in Figure 1.1, where Holmes is the row player (choosing the row), and Moriarty the column player. What option will the players choose? Actually neither option for either player

	C	D
C	$0, 2$	$2, 1$
D	$1, 0$	$0, 3$

Figure 1.1: The game Holmes and Moriarty are playing.

would be ruled out by any of the 'optimality' operators we consider later; intuitively this is because none of the options are obviously irrational. Indeed this is an example where a one-shot analysis of the situation does not have much to say. Holmes' famously flawless and yet insightful logic should reveal to him what is the best solution in his dilemma: alight at Canterbury or stay on until Dover. Let us imagine a dialogue between Holmes and Watson[5].

> **Watson:** Moriarty wants to go to Dover, therefore you should alight at Canterbury, and live to capture that swine on a later date.
> **Holmes:** How simple-minded you are sometimes Watson. Apparently you have forgotten with whom we are dealing. Do you really think that Moriarty is not able to put himself in my shoes, and to reason in precisely that way?
> **W:** Oh yes I see Holmes, so you mean you should alight at Dover, because Moriarty knows that you know that he wants to go there, and so will expect you to alight at Canterbury, and so will alight there to try to catch you. How clever you are, to out-think him that way.
> **H:** Again Watson you are not thinking enough. Moriarty will be able to perform that reasoning as well...
> **W:** So you mean I was correct before, but for the wrong reasons: alight at Canterbury, also that way you can be back in time for tea!
> **H:** (Sighs) I fear you are not getting my point.

What is Holmes' point? The fact is that if there were some deductive reasoning that could lead Holmes to see that C (or D) was the best option, then since Holmes' opponent is also highly intelligent, *Moriarty could also follow that reasoning*, thereby alighting at C (respectively D), and shooting Holmes. In which case clearly Holmes' reasoning did not in fact lead him to the best option.

[5] We mercilessly misrepresent the characters of Holmes and Watson to fit stereotype rather than their actual nature in the books by Conan Doyle. Also, for the sake of our story, recall that the cunning detective's trusty side-kick is not present.

Therefore since Holmes cannot make a decision based purely on deduction, he is forced to throw reason to the wind and alight wherever his intuition tells him to. I.e., he cannot actually make the decision, but, *if we assume that Moriarty would be capable of re-creating Holmes' reasoning*, the best option for Holmes appears to be that he must randomise between the two options.

That is, Holmes should play a so-called ***mixed strategy***. In the case of games with *ordinal* preferences, mixed strategies would simply be *sets* of strategies; in this case since there are only two options, $\{D, C\}$ is Holmes' mixed strategy. Mixed strategy profiles, that specify a mixed strategy for each player $i \in N$, are in this case just *subgames* of the original game. There would be several different ways to lift the existing preference relation over pure strategy profiles to a relation over mixed strategy profiles. It is not clear what grounds to use to choose among the different liftings, but we do not pursue this matter further, since mixed strategies are generally only considered in the case of cardinal preferences, where a much more fine-grained distinction between mixed strategies is available. Indeed the literature generally only considers mixed strategies in terms of cardinal utility (see for example [Osborne and Rubinstein, 1994, Chapter 3]), so we will take 'mixed strategy' to imply that the underlying game is one of cardinal utility.

For games with cardinal preferences, mixed strategies are more complicated entities: in these games, a mixed strategy for player i is a *probability distribution over* T_i. (A probability distribution over T_i, for *finite* T_i, is a function $\sigma : T_i \to [0, 1]$ such that $\sum_{s_i \in T_i} \sigma(s_i) = 1$.) There is a debate in the literature of game theory as to how to interpret mixed strategies. We favour taking them simply to mean that the player literally randomises over his choices with the relevant probabilities, but this only colours the way we talk about them, and not the content of any theorems we prove that relate to them.

Another common interpretation has it that mixed strategies really represent a *belief* by the opponents about how a player will play. This is argued for in for example [Aumann and Brandenburger, 1995]. While we recognise that one can represent some elements of a player's beliefs as a mixed strategy, still to say that the mixed strategy that a player actually players *is* a belief by the opponents is a superficial treatment of the notion of belief. If a player i plays a mixed strategy σ, it would mean that all of the players have the *same* belief regarding i's behaviour. More importantly, such a simplistic approach means that any kind of higher-order belief (i's belief about j's belief) collapses. This, as we indicate in Chapter 4, might well make sense in an *equilibrium where all beliefs become common belief*. So it arguably fits with some steady-state interpretation of game theory (and it is indeed Nash equilibrium, and so a steady-state interpretation, that is considered in [Aumann and Brandenburger, 1995]). However, clearly it does not suit the *deductive* approach that should be applied to truly one-shot interaction situations.

To repeat: in the deductive approach, playing a mixed strategy should really mean randomising, with the allotted probabilities, between the different options, and does not represent a belief or 'conjecture' by the other players about what the one player will

do. We do not want always to assume that the players have this option to randomise. Indeed, if this option is available to the players then we might want to say that really they are in a bigger game, the so-called '*mixed extension*' of the original game, in which the strategies are the mixed strategies from the original, and the payoffs are given by what is called the "expected utility" function.

For *finite* S_i, we write ΔS_i to mean the set of probability distributions over S_i. Then in a game with cardinal preferences the set of mixed strategies of player i is ΔT_i, Generally we will use σ to refer to mixed strategy profiles (yielding a mixed strategy $\sigma_i \in \Delta T_i$ for each player i). The canonical way to extend a utility function over pure strategy profiles to a utility function over mixed strategy profiles, that defines the preferences in the mixed extension, uses the notion of "expected utility" (that might better just be called "mixed utility"):

$$\mu_i(\sigma) = \sum_{s \in T} \sigma(s) \cdot \pi_i(s).$$

Note that we cannot without some other stipulation extend this definition to the case where T is infinite. That is because examples can be constructed in which the expected utility of a given strategy profile would be infinite. So we will assume in general, when we talk about expected utility and mixed strategies, that there are a finite set of strategies. (Another solution is to consider only probability distributions with *finite support*, i.e. in which only a finite number of strategies are assigned non-zero probability, or to place some restriction that excludes 'badly-behaved' utility functions, but we do not need to go into any further detail here.)

Pure strategies are in effect simply 'degenerate' cases of mixed strategies, in which all of the probability mass is assigned to a single element, and sometimes we will write a term s_i denoting a pure strategy to mean the corresponding mixed strategy that associates probability 1 to s_i, and 0 to all other of i's pure strategies. (So in set-theoretic notation it would be the function $\{(s_i, 1)\} \cup ((T_i - \{s_i\}) \times \{0\})$).

The next concept we need to formalise in this Section is that of game reduction operators, or 'optimality operators'. An optimality operator for player i is supposed to say which strategies i should 'throw out' on the grounds that they are sub-optimal. We want our approach to be as generic as possible, and so while many specific optimality operators exist, the results we will present in this Chapter will hold for optimality operators that satisfy a certain condition of *monotonicity*. An individual optimality operator[6] for player i takes a game and a restriction and returns a set of strategies.

Definition 1.2. An ***individual optimality operators*** for player i is any (class-)function that, given a game $G = (T_j, \geq_j)_{j \in N}$ with $i \in N$ and a restriction $S \subseteq T$, returns a set $S_i' \subseteq T_i$ of i's strategies.

[6]We sometimes omit the word 'individual', which is there to distinguish it from 'collective' optimality operators that we introduce shortly.

Often it will be convenient to fix a game G and to consider optimality operators as functions $O_i^G : 2^T \to 2^{T_i}$. (And sometimes when G is clear from the context we will drop the superscript G.)

This definition is of course a little too abstract to really capture any notion of optimality. In actual examples optimality operators will be defined in terms of the players' preferences, and will capture the notion that optimal choices are in some sense *preferred* over sub-optimal choices. We will give examples of such optimality operators later in this Section, but first let us give two important properties of optimality operators that we will consider:

An operator O is ***contracting*** if for all restrictions S, $O(S) \subseteq S$. And O is ***monotonic*** if for all restrictions S and S', if $S \subseteq S'$ then $O_i(S) \subseteq O_i(S')$.

The idea behind the argument given to the individual optimality operator for a particular game is that it is intended to represent the restriction of the game that the player thinks she is actually playing in. This will become more formal when we introduce belief models in Section 1.3 to capture the idea of a player 'thinking' (or rather, 'believing') something. For now the operators remain purely algorithmic, or procedural. We will be interested in combining them and in iterating the resulting operator, that we will call a ***collective optimality operator***, or sometimes (again) just an 'optimality operator'. So given a family of individual optimality operators $(O_i)_{i \in N}$, let O denote the operator from restrictions to restrictions, i.e. $O : 2^T \to 2^T$, defined as follows:

$$O(S) = \prod_{i \in N} O_i(S).$$

Clearly if each O_i is contracting or monotonic then O is contracting or, respectively, monotonic with respect to the component-wise subset ordering.

Optimality operators actually operate only on rectangular restrictions, and collective operators return rectangular restrictions. So we could have defined them in terms of subgames, as for example in [Apt, 2007c]. We prefer the more general formulation in terms of restrictions simply because it fits better with the rest of our notation.

Fixing some game G, we will be interested in iterations of this collective operator, starting with the largest restriction, that corresponds to the initial game. Let $\mathcal{ON}$ denote the class of all ordinals; then given some game $G = (T, \geq)$ and an optimality operator O for $\alpha \in \mathcal{ON}$, O^α is the operation corresponding to α applications of O. Precisely, it is defined as follows, where (and this is a convention we maintain throughout) β is an arbitrary ordinal and λ a limit ordinal:

$$\begin{array}{ll} O^0(T) & = T \\ O^{\beta+1}(T) & = O(O^\beta(T)) \\ O^\lambda(T) & = \bigcap_{\beta<\lambda} O^\beta(T) \end{array}$$

To make the notation more elegant, we often, when it is clear from the context what T in question, write simply O^α for $O^\alpha(T)$.

A restriction S is a ***fixpoint*** of O if $O(S) = S$. Assume fixed some game $G = (T, <)$. Then for $\alpha \in \mathcal{ON}$ with $O^\alpha(T)$ a fixpoint, we call that (obviously unique) $O^\alpha(T)$ the ***outcome*** of O on G, and denote it by O^∞_G.[7]

Definition 1.3. If $G = (T, <)$, then we call the least ordinal α such that $O^\alpha(T) = O^\infty_G$ the ***O-outcome ordinal*** of G, and write it α^O_G.

For each ordinal α and optimality operator O, the result of iterating O α times is a '*solution concept*', to use the terminology of game theory. The most commonly-considered such solution concept, for a given optimality operator, is its outcome. In Section 1.3 we will be interested in the outcomes of (collective) optimality operators, and will make appeal to Fact 1.1.2.

Fact 1.1. *We are guaranteed that, for any game G:*

1. *If O is contracting then it has an outcome (because T is a set).*
2. *If O is monotonic then it has an outcome, which is the* largest fixpoint $\bigcup\{S \subseteq T \mid S \subseteq O(S)\}$ *(immediate corollary of [Tarski, 1955, Theorem 1]).*

Note that while if an operator is contracting it need not be monotonic, or vice-versa, still the *outcome* of a monotonic operator and its 'contracting version' coincide, in the sense that given some monotonic operator O, and defining $\overline{O}(S) = S \cap O(S)$ as its *contracting version*, we have the following Fact (cf. [Apt, 2007b, Note 1]).

Fact 1.2. *For any $\alpha \in \mathcal{ON}$, $O^\alpha = \overline{O}^\alpha$.*

We can now look at some particular instances of optimality operators from the game-theoretical literature, where each different optimality notion has several different specific instantiations.

The first group of operators that we will look at are those induced by the elimination of strictly dominated strategies. A strategy s_i is ***strictly dominated*** by a strategy s'_i in the context of S_{-i} if

$$\forall s_{-i} \in S_{-i},\ (s'_i, s_{-i}) >_i (s_i, s_{-i}).$$

(For cardinal utility and mixed strategies, simply replace, here and in the rest of the Chapter, $(s'_i, s_{-i}) >_i (s_i, s_{-i})$ by $\mu_i(s'_i, s_{-i}) > \mu_i(s_i, s_{-i})$.) We write $nsd_i(s_i, s'_i, S_{-i})$ to mean that s_i is *not* strictly dominated by s'_i in the context of S_{-i}. Now there are several ways in which we can use this property to induce an operator, which will all have the following form, with different instantiations for A and B:

$$O^G_i(S) = \{s_i \in A \mid \forall s'_i \in B,\ nsd_i(s_i, s'_i, S_{-i})\}.$$

[7] We thus use the same word 'outcome' for both the outcomes of a game and the outcome of the iterated elimination of non-optimal strategies. We do not anticipate that this will cause any confusion, but note of course that the outcome of iterated elimination will in general not yield a *single* outcome in the other sense.

Now there appear to be eight different versions of this operator given by instantiating A with either S_i or T_i, and B with one of S_i, T_i, ΔS_i or ΔT_i (where, as we assume throughout, $G = (T_i, \geq_i)_{i \in N}$, or $(T_i, \pi_i)_{i \in N}$ if it is a game with cardinal preferences). Some of these operators coincide, but let us use them to introduce our terminology, that we will use for all of the different optimality notions.

If A is instantiated with S_i, we call the operator the ***contracting*** form (since clearly then it will be contracting), and if A is instead instantiated with T_i we call it the non-contracting form (even if it might in fact happen to be contracting). When B is either S_i or ΔS_i then we call the resulting operator the ***local*** form, and otherwise we call it the ***global*** form. The idea behind the local form will be that player looks only at the possibilities given by her information about the game in order to determine whether there is a better strategy to play, and in the global version does not 'forget' possibilities that might be better. The distinction between global and local (and contracting and non-contracting) optimality operators is due to [Apt, 2007b; 2007c]. The last piece of terminology is not surprising: if B is either ΔS_i or ΔT_i then we talk about the ***mixed*** form, otherwise we talk about the ***pure*** form (or just drop the qualifier altogether).

The observation that the contracting form is contracting is obvious, and indeed clearly holds no matter what property we would put for nsd_i. That the *global* form is *monotonic* is only a little less obvious, but essentially just uses the fact that it can be defined by a formula which is positive in the argument the operator takes.

Our main interest in this Chapter will be in the *global* forms of the operators, because in general only the global forms are monotonic. Here and in what follows, by Fact 1.2, as long as we are only interested in iterations of a monotonic operator from the initial game, then it does not matter whether we consider the contracting or non-contracting forms.

Furthermore, since we are considering only these iterations starting from the initial game, then notice that on *finite* games (i.e. in which each player's strategy set T_i is finite), there are actually only two operators. That is established by [Apt, 2007c, Lemma 2], which implies that on finite games the local and global forms of each of the considered operators coincide. Therefore the only distinction that remains is between the mixed forms. A standard example (cf. [Osborne and Rubinstein, 1994, Figure 61.1]) shows the pure and mixed forms do not coincide; see the game illustrated in Figure 1.2. (We read this as a game with cardinal utilities, because there is no way to make

	L	M	R
U	$1, 3$	$1, 0$	$0, 1$
D	$0, 0$	$0, 3$	$3, 1$

Figure 1.2: A finite game distinguishing pure and mixed strict dominance.

the distinction with ordinal mixed strategies) In that game, R for the column player, b, is not strictly dominated in the context of $\{U, D\}$ by either of the *pure* strategies L or

M, but *is* strictly dominated in the context of $\{U, D\}$ by for example the mixed strategy σ_b with $\sigma_b(L) = \sigma_b(M) = 0.5$, since then for any of the row player a's strategies $s_a \in \{U, D\}$, we have $\mu_b(\sigma_b, s_a) > \mu_b(s_b, s_a)$.

Notice that this example illustrates that neither player actually has to end up *playing* a mixed strategy for the *possibility* that they *could* play a mixed strategy to affect the outcome. Specifically, if we remove R (because it is strictly dominated by $\{(L, 0.5), (M, 0.5)\}$), then D becomes sub-optimal for the row player, since the opponent is going to play in $\{L, M\}$, so we can remove D, and then go on to remove M, since in the context of $\{U\}$ it is the only undominated strategy for the column player. So then in the outcome, neither player has a choice left, so they play L and U, both pure strategies. However, the outcome of eliminating strategies dominated by a *pure* strategy is the entire game, since no strategy is dominated.

Conditions for the iterated elimination of strictly dominated strategies on finite games, in terms of the beliefs (actually they also considered knowledge) were first given in [Tan and Werlang, 1988], and their result is one that we will generalise below.

The notion of weak dominance is a refinement of strict dominance: a strategy s_i can be weakly dominated by s_i' in the context of S_{-i} even if for some $s_{-i} \in S_{-i}$, s_i' does not do strictly better against s_i, *as long as it never does worse*. The formal definition, or rather the schema that defines the same forms as in the case of strict dominance, is as follows: s_i is ***weakly dominated*** by s_i' w.r.t S_{-i}, denoted $wd_i(s_i, s_i', S_{-i})$, if:

$$\begin{aligned} & \forall s_{-i} \in S_{-i}, (s_i', s_{-i}) \geq_i (s_i, s_{-i}) \\ \text{and} \quad & \exists s_{-i} \in S_{-i}, (s_i', s_{-i}) >_i (s_i, s_{-i}). \end{aligned}$$

Now, although weak dominance has *prima facie* intuitive appeal, it turns out to be a rather less mathematically well-behaved notion than strict dominance. The first point to notice, that disqualifies it from the scope of the theorems we will prove in this Chapter, is that *neither its local nor global forms are monotonic*. For instance in the game depicted in Figure 1.3, D is not weakly dominated in the context of $\{L\}$, yet in the context of the larger set $\{L, R\}$, it *is* weakly dominated. It is precisely examples

	L	R
U	$0, 0$	$1, 0$
D	$0, 1$	$0, 0$

Figure 1.3: A game illustrating the non-monotonicity of weak dominance.

like this that are ruled out by the monotonicity of an operator, so such examples do not exist for strict dominance.

In this example, the outcome is $(\{U\}, \{L\})$. Notice though that intuitively speaking it is not clear why the players would play these choices, given that together they yield a least preferred option (for both players). In terms of beliefs, the justification should be that although e.g. the row player *believes* the column player will play L, still she

should leave open the possibility that he will play R. That is not something that the epistemic framework of this Chapter can deal with in general, and it is a topic that we will return to in the next Chapter, when we look at counterfactual beliefs.

Although we will not consider this next issue in any detail, notice also that there would be different ways of putting together an operator for weak dominance that would yield different results. What we mean is that although we have defined collective optimality operators as the intersection of individual optimality operators, there are other possible ways of doing this. Taking an intersection effectively means applying the individual operators simultaneously, whereas we might want instead to *iterate* the individual operators. Then in our example, we might apply first the operator for the row player, thus obtaining (since D is weakly dominated by U in the context of $\{L, R\}$) the restriction $(\{U\}, \{L, R\})$, and *only then* apply the operator for the column player, which in this case will leave both strategies, so that the outcome of such an operator will be $(\{U\}, \{L, R\})$, which is clearly a different outcome from the simultaneous version.

One could even combine the individual optimality operators in more ways, by only *partially* applying the individual operators. So, to take an example directly from [Osborne and Rubinstein, 1994, Figure 63.1], in the game depicted in Figure 1.4, M and R are both weakly dominated by L. If we first remove R, U becomes weakly dominated leaving the outcome $(\{D\}, \{L, M\})$, but if we instead first remove M, then D is dominated, leaving the disjoint outcome $(\{U\}, \{L, R\})$. Such a situation can never arise in the case of the monotonic operators.

	L	M	R
U	$1, 1$	$0, 0$	$1, 1$
D	$1, 2$	$1, 2$	$0, 0$

Figure 1.4: A game illustrating the order-dependence of individual weak dominance operators.

The final kind of optimality we consider is a different strengthening of strict dominance: best response. The idea here is that a strategy is only optimal if it can be justified ('rationalised', to borrow the terminology of Bernheim [1984] or Pearce [1984]) by a belief that the other players will play in such-and-such a way. Believing that players will play in such-and-such a way could mean two things: thinking that they will play according to a given pure strategy; or thinking they will play according to a given mixed strategy.

In the former case, we say that s_i is a *point best response* in the context of S_{-i} and among B (where as above $B \in \{S_i, T_i\}$ determines whether it is the local/global and pure/mixed form of the property), written $pbr(s_i, B, S_{-i})$, just when

$$\exists s_{-i} \in S_{-i} : \forall s_i' \in B(s_i, s_{-i}),\ \geq_i (s_i', s_{-i}).$$

As observed in [Benthem, 2007b], this is just a quantifier permutation of the condition of not being strictly dominated: overloading notation we could write $nsd(s_i, B, S_{-i})$ to mean that s_i is not strictly dominated by any strategy in B in the context of S_{-i}, which would express the following condition:

$$\forall s_i' \in B \exists s_{-i} \in S_{-i} : (s_i, s_{-i}) \geq_i (s_i', s_{-i}).$$

So clearly no best response is strictly dominated, but the converse does not in general hold; witness Figure 1.5 (cf. [Benthem, 2007b, Proposition 3]). There D is never a

	W	M	D
Y	$1, 2$	$1, 0$	$1, 1$
N	$0, 0$	$0, 2$	$2, 1$

Figure 1.5: A game illustrating the difference between strict dominance and point never best response.

best response to a *point* belief, but it is certainly not strictly dominated by any (pure or mixed) strategy. To see that D is never a best response to a point belief, notice that there are only two point beliefs concerning the row player's strategy that we could assign to the column player: Y and N. And in either case, rationality would dictate that the player not play D; a belief in Y would mean the column player had better play W, and a belief in N would mean the column player had better play M.

The '*point*' case is a bit restrictive, in the sense that it is not necessarily very realistic to assume that players have a definite idea of what the other players will do: indeed, in the game in Figure 1.5, it seems reasonable to think that the column player, being on the face of it unsure whether the row player will play Y or N, might well play D, in order to 'hedge her bets'. And indeed, in the cardinal utility case we can see that playing D *is* now a best response, to $\{(Y, 0.5), (N, 0.5)\}$.

There are two ways to define a mixed strategy profile of one's opponents. One can take ΔS_{-i}, the set of probability distributions over S_{-i}, but notice that this assumes some form of *coordination or correlation* amongst the players, which of course goes against the spirit of this view of strategic games as representing a specific kind of one-shot interaction (cf. [Bernheim, 1984, page 1014]). So it is only reasonable to demand that i play best responses to a subset of these probability distributions, namely those in which there is no coordination. These can be represented by the product $\prod_{j \in N - \{i\}} \Delta S_j$. We denote this set of uncoordinated strategy profiles of i's opponents as $\Delta_u S_{-i}$. (Of course for 2-player games the two coincide.)

Thus we can say that a strategy is a ***best response*** (to an uncorrelated mixed strategy of the opponents), written $ubr(s_i, B, S_{-i})$, if $pbr(s_i, B, \Delta_u S_{-i})$, and that it is a ***correlated best response***, written $cbr(s_i, B, S_{-i})$ when $pbr(s_i, B, \Delta S_{-i})$. [Pearce, 1984, Lemma 3] shows that $nsd(s_i, S_i, S_{-i})$ if $cbr(s_i, S_i, S_{-i})$ for the case of finite games. That is: if the game is finite, a correlated best response is the same thing as a strategy

that is not strictly dominated. Furthermore, it is a consequence of [Apt, 2007b, Theorem 3] that iterations of the local and global versions of those operators coincide on finite games. There is no distinction to be made between any of those operators on finite games. (That Lemma of Pearce is generalised to cover a certain class of infinite games in [Zimper, 2005].)

Let us remark that mixed strategies here *do* take a kind of belief interpretation, and are sometimes talked about as 'beliefs' in the relevant literature. This is not the same kind of steady-state interpretation as the kind that we mentioned above when talking of mixed strategies, that reduces all beliefs to common beliefs. That is: it does not require that all players refer to the *same* mixed strategy profile of their opponents. Nonetheless, to restrict the description of the belief state of a player to just a (correlated or uncorrelated) strategy profile of her opponents is still to sell short the possibilities of epistemic analysis. Modern-day interactive epistemology uses more detailed models including relational models [Fagin *et al.*, 1995], Harsanyi type spaces [Harsanyi, 1968], Aumann structures [Aumann, 1976], (which are instances of modal logic models) and neighbourhood models [Heifetz, 1996]. These various different kinds of models give a richer account of what a belief state is. They all share the feature that they can be 'unfolded' to make statements *recursively*, so that they express not just whether player i believes something, but also whether player j believes that player i believes it, and so on. This kind of 'higher-order' belief – belief about belief – is crucial to understanding many social (interactive) situations.

Back to the question that will be most relevant for the rest of the material of this Chapter: are any of the operators induced by the notion of best response *monotonic*? The answer, which suits our interests, is that once again the global versions are indeed monotonic. This, as with the case of strict dominance, is explainable just by looking at the logical (quantifier) form of the properties: they are all positive in the relevant argument.

Let us briefly recall then that we have examined three different examples of what can be meant by 'optimality', and have illustrated several distinctions that can be made within these different examples. The optimality operator approach abstracts away from all of this however, and allows us to reason about all the operators that are, for example, contracting and monotonic. We have seen that there are several of these: strict dominance by a pure or mixed strategy; and three forms of best response (one of which collapses in most cases to strict dominance by a mixed strategy).

1.2 Heuristic treatment

In this Section we will take a brief tour and summary of the technical material and results that we will present in full detail in Sections 1.3 and 1.4.

In the rest of this Chapter we restrict attention to that class of game reduction procedures that satisfy the property of *monotonicity*, and provide a unified foundation for them in terms of beliefs about rationality. The results that we prove generalise stan-

dard results from the literature. The first sources for such results are [Bernheim, 1984; Pearce, 1984; Tan and Werlang, 1988]. However, our framework, when it involves so-called 'relational models' for beliefs (i.e. in Section 1.3), will be closer to that in the survey paper [Battigalli and Bonanno, 1999].

Let O denote some monotonic optimality operator, and G denote a game. Then the first result that we prove will be the following two theorems, which are respectively similar to (but, as we will explain, generalisations of) Propositions 3.10 and 3.11 from [Battigalli and Bonanno, 1999].

Theorem 1.1. Common true belief in O-rationality entails players in G only choosing strategy profiles in O^{∞}_{G}.

Theorem 1.2. There is a model of G in which if the players choose strategies in O^{∞}_{G} then there is common true belief of O-rationality.

These theorems are together sometimes taken to mean that common true belief in O-rationality is *equivalent* to not playing strategies that survive the iterated elimination of non-O strategies (i.e. to O^{∞}_{G}). Let us not shy away from emphasising that while one direction of the implication is indeed given by Theorem 1.1, the degree to which the second theorem indicates an *equivalence* is unclear. The line taken in [Battigalli and Bonanno, 1999] is that the equivalence *is* established, but that it is "made even more transparent within a universal type space, which – by definition – contains all the conceivable hierarchies of beliefs" (op.cit. p.14 n.36). In Chapter 2 of this Thesis we will take a modal logic perspective on universal type spaces, but let us pre-emptively say that we do not find that universal type spaces allow a better statement of the equivalence, since mathematically speaking nothing is added, and conceptually we do not find any clarity in the notion of a universal type space. Still, we will accept that these theorems do establish something close to an equivalence, and certainly do not yet have a better proposal of our own, so continue to call what they establish an 'equivalence'.[8]

The equivalence that we prove extends the standard result by covering various different kinds of optimality notion. Our equivalence theorems will cover *arbitrary* games, and not just finite games, as was the case in previous statements found in the literature along the lines of these theorems.

In giving the most general form of Theorem 1.1, we will be interested to use the smallest number of assumptions about beliefs (so, to borrow a term from the next Chapter, the *weakest* (smallest) possible 'logics' of belief). Conversely, in proving the most general form of Theorem 1.2, we will want to consider the *smallest* classes of models of belief, where a large number of properties are respected (so the *strongest* logics).

[8] Actually, we establish something closer to an equivalence than is established by [Battigalli and Bonanno, 1999, Proposition 3.11], since that says 'for every strategy such that... there is a model such that ...', whereas we say 'there is a model such that for every strategy such that ...'. We do not make much of our minor strengthening of the result here, but do return to this $\forall\exists$ to $\exists\forall$ quantifier shift below in discussion with the later theorems where it is a little more substantial.

One kind of property that we will look at, in generalising Theorem 1.1, are the 'introspection' properties sometimes attributed to players. These say that if a player believes something then she believes that she believes it (so-called positive introspection), and if she does not believe something then she believes that she does not believe it (negative introspection). Partition structures [Aumann, 1976] for example entail that players' true beliefs are *positively* and *negatively introspective*. In fact in modal logic terminology they are 'S5' models, meaning that there is positive and negative introspection, and also that beliefs are always true (indeed Aumann talks about "knowledge"), and finally that the players believe the consequences of their beliefs (so if they believe E and that E implies F, then they believe F). We show that these properties of players' beliefs are *not* needed in order to establish Theorem 1.1. This result is thus in the spirit of [Samet, 1990], who generalises to so-called 'S4' models in which negative introspection can fail, the result for partitional S5 models in which negative introspection *cannot* fail, established in [Aumann, 1976] (the so-called 'agreement theorem'[9]). We show that negative introspection is not needed, and nor are positive introspection or correctness of beliefs.

We also look, in the last Section, at whether believing the consequences of one's beliefs is needed in order to establish a version of the result. On all relational models, players believe the consequences of their beliefs, but this is not so on some so-called 'neighbourhood models', of which relational models are special cases. In neighbourhood models, one simply *lists*, for each state, the propositions that each player believes at that state. These lists do not make any assumptions about the way in which the player has put those pieces of information together. Neighbourhood models are very permissive, in the sense that they make very few assumptions about the properties of the players' beliefs. So neighbourhood models do not in general require introspectivity, but more significantly, in a neighbourhood model a player can believe an event E, believe that E entails F, and still not believe F. Neighbourhood models are studied in [Chellas, 1980; Hansen *et al.*, 2009]. and are considered as models for beliefs in [Lismont and Mongin, 1994; Heifetz, 1996; 1999].

We prove Theorem 1.1 with respect to arbitrary relational models. But in the next Section we also prove two forms of the same result for neighbourhood models. Conversely, we state Theorem 1.2, that states the *existence* of a model, in its strongest form, so for the most restrictive class of models, S5 (partitional) models.

We will also see (Fact 1.7) that for every ordinal α there are games that require precisely α rounds of elimination of strictly dominated strategies before there are no longer any dominated strategies.

This motivates us to look more closely at the connection between these transfinite eliminations and transfinite levels of mutual belief of rationality, and to ask the question

[9]That theorem is not directly related to any of the work we present here; it says that, in a probabilistic context, if players have a common prior and common knowledge of their posterior beliefs (i.e. the relativisation of their prior to the information represented by the relational structure), then the posterior beliefs are in fact the same. Different non-probabilistic versions are presented in [Samet, 2006] and [Dégremont and Roy, 2009].

about intermediate stages, *before* full common belief is reached. We credited [Tan and Werlang, 1988] above with establishing an equivalence between common true belief of rationality for the case of strict dominance in finite games. They actually did more, proving a result about intermediate stages (op.cit., Theorems 5.1 and 5.3). So they have the following form:

Theorem ([Tan and Werlang, 1988])**.** *Mutual true belief of order m in NSD-rationality is equivalent to $m+1$ rounds of elimination of strictly dominated strategies.*

(Here 'NSD-rationality' means not playing strategies that you believe to be strictly dominated.)

If we want to generalise this statement to the infinite case, it turns out that there is a problem. When there is mutual belief of order ω_0 in A, we say that there is ***finitary common belief*** of A. and when there is mutual belief of order α of A *for any infinite ordinal* α, this we call ***absolute common belief*** of A, or just *common belief* of A. In the standard, relational, models for belief (of which, recall, S5 partition structures are a special case), finitary common belief *is the same thing as* absolute common belief. Relational models are therefore not adequate for reasoning about transfinite beliefs, including about transfinite belief in rationality.

That observation has been made in a number of places, for example in [Barwise, 1988; Heifetz, 1999; Benthem and Sarenac, 2004]. Barwise indicates the difference between the 'iterative' and the 'fixpoint' definitions of common knowledge, which in our terminology correspond to *finitary* and *absolute* common belief, and shows that in relational models they coincide but that in so-called "situation semantics" they do not. Situation semantics turns out to be essentially a notational variant of neighbourhood semantics; this point is made formal in [Lismont, 1994]. Heifetz [1999] studies infinitary axiom systems for reasoning about common belief and shows their completeness with respect to monotonic neighbourhood models.[10] This is then used to establish the difference between the finitary and absolute forms of common belief in neighbourhood semantics. In [Benthem and Sarenac, 2004], the authors consider a special case of neighbourhood models, that uses a topology to represent the information of players, and show that there too the finitary and absolute versions of common belief can be separated.

In Section 1.4, we show that neighbourhood models are adequate for this kind of transfinite reasoning about games. In the context of neighbourhood models, assuming only that the players' beliefs respect a certain kind of introspection property, we prove Theorem 1.4.

Theorem 1.4. Mutual true belief of order α in O-rationality entails only playing strategies that survive $1+\alpha$ rounds of elimination of non-O strategies.

[10] We postpone discussing axiom systems and completeness until the next Chapter, but roughly speaking: an axiom system is complete with respect to a class of models if every sentence it can prove is true in every model, and vice-versa.

(Note that we write $1+\alpha$ because for *finite* ordinals α, there is one round of elimination 'for free', whereas this is not the case for infinite ordinals. That is to say: rationality *by itself* entails not playing strategies that are eliminated immediately (in one round), so for the case where $\alpha = 0$ we clearly get $\alpha + 1$ rounds of elimination. And (as we prove formally below) this continues, so that rationality combined with m-level belief in rationality entails $m + 1$ rounds of elimination, *until* we hit the first infinite ordinal ω_0, where rationality plus ω_0-level mutual belief in rationality entails ω_0 rounds of elimination of non-optimal strategies, and so on.)

We also prove Theorem 1.3, which is a slight variant of Theorem 1.4 that does *not* require the mentioned introspection property. For that we use a notion of mutual belief that we call 'co-mutual belief'. Co-mutual belief is equivalent to mutual belief on the standard relational models, but diverges from it on neighbourhood semantics.

There is additionally a 'converse' direction to Theorem 1.4:

Theorem 1.5. There is a model of G in which for all ordinals α, the players choose strategies in $O_G^{1+\alpha}$ iff there is α-level mutual true belief of O-rationality.

Indeed, it is *this* Theorem that does not in general hold for relational models, and so it is here that we have recourse to neighbourhood models. We take these theorems to mean that, as far as the framework of this chapter allows us to establish it, *in the relevant class of models*, α-level mutual true belief of rationality is equivalent to $1 + \alpha$ rounds of iteration of non-optimal strategies. Theorem 1.5 does not hold for transfinite ordinals α if we restrict our attention to relational models.

As we have said, neighbourhood models are very permissive. *Topological* models are less permissive, and in them all players are *positively introspective* regarding their beliefs, and, as in relational models, in topological models players *do* always believe the (things they believe to be the) consequences of their beliefs. As with Theorems 1.1 and 1.2, for generality we prove Theorems 1.3 and 1.4 in the most permissive cases of neighbourhood models. And Theorem 1.5 we prove with respect to topological models.

The extent to which Theorems 1.4 and 1.5 really establish an equivalence is, as in the case of common belief, not entirely clear. Nonetheless, we do state Theorem 1.5 in a stronger form than it is usually stated in the literature: that there is a model such that for every ordinal, for every strategy... We can call this the '$\exists\forall\forall$' formulation. In the formulation (of the finite version) in the literature, it is stated in the strictly weaker '$\forall\forall\exists$' version: that for every n, for every strategy... there is a model such that...[11] We prefer our formulation. Mainly this is because mathematically speaking it is a strictly stronger result, i.e. the model no longer depends on the ordinal or the strategy. However, it also has the advantage of some conceptual appeal: in one model we are able in some sense to 'rationalise' every possible play of the game. Yet we will show that, even in the finite case, although the $\forall\forall\exists$ version continues to hold, if we assume that players are *negatively introspective* then the $\exists\forall\forall$ version does *not* hold.

[11] We make these elided phrases clearer below once we have introduce the relevant technical notation!

1.3 Common belief in rationality

It is time to make formal our talk of beliefs and belief models. We will start in this Section by presenting formal definitions of the standard relational models for beliefs, that we have already mentioned informally above. These have, since the work of Kanger [1957] and Kripke [1959], been standard in modal logic and partitional models, developed independently by Aumann for reasoning about knowledge and common knowledge [Aumann, 1976], are a special case of relational models. We do not yet introduce a formal language, and so do not make any distinction between syntax and semantics, postponing that until Chapter 2. (Aumann in his early work does not make such a distinction, but in more recent work [Aumann, 1999] favours an approach that does distinguish between syntax and semantics.)

[Aumann, 1976] formalises a notion of *common knowledge*. Recall that we prefer to talk of 'common belief'; common knowledge can be seen as a special case of this, in which beliefs are *never* incorrect. This is partly because we will be considering more general classes of models than the partitional spaces, and in these more general classes there are models in which the modalities represented do not have properties that we would want to insist that knowledge have. (Or at least one property that should distinguish knowledge from belief: that the former cannot be incorrect.)

We will define the introspection properties of beliefs, then look at levels of mutual belief, see the two definitions of common belief, and that they coincide on relational models. We also present our definition of 'co-mutual belief', which is equivalent, on relational models, to mutual belief, but as we will see in the next Section diverges in the case of neighbourhood models (that we introduce there). Another essential ingredient will of course be the definition of *rationality*.

All of the different models we will look at are based on a 'state space', a set of 'states', that might also be called 'possible worlds'. A state or possible world specifies which of the 'relevant' non-epistemic properties hold. In the context of a game, we take the relevant non-epistemic properties to be just the choices made by the players. The context, i.e. the model in which a state resides, in turn provides the epistemic properties (the properties of the beliefs of the players). In Chapter 3 we will introduce more sophisticated models, essentially those from [Board, 2002; Baltag and Smets, 2006], to represent the beliefs of players. These will have the capacity to represent *conditional* beliefs, which game-theoretical considerations show to be worthy of study, for they, as we argue there and in Chapter 4, are necessary for a correct understanding of the epistemic analysis of non-monotonic optimality operators, and the reasoning about *counterfactuals* implicit in the so-called 'many-moment interpretation' of extensive games that we will adopt in that Chapter [Stalnaker, 1996; Bruin, 2004]. Counterfactuals quite simply don't matter in what we consider in the present Chapter, where we focus on an analysis of possible justifications, in epistemic terms, for playing within the sets generated by iterations of monotonic operators. Therefore we will present simpler *unconditional* belief models in this Chapter.

In fact it will turn out that everything we need to do in this Chapter can be done

using a rather restricted notion of state space. Suppose that we did not consider *arbitrary* state spaces, but said that, for whatever game was being modelled, the state space *is* the set of strategy profiles of that game. In fact all of the results would still hold if we were to do this, and the only reason we avoid doing so is to avoid at the same time the potential charge of over-simplification. Indeed, some of the results that we present (for the record: Theorems 1.1, 1.3 and 1.4) are strictly stronger as stated in this general form. Still, let us emphasise that the alternative approach, of letting the state space be the space of strategy profiles, would arguably be consistent with the one-shot deductive interpretation of game theory that we have in mind. (Another, more frivolous, argument in its favour is that it would simplify notation!)

So we consider arbitrary state spaces, but not without reservations. Subsets of the state space W, i.e. elements of 2^W, are called ***events***, or sometimes ***propositions***. If $u \in E$ we say that E is 'true' at u. (Later, in Chapter 2, when we introduce a distinction between the syntax and semantics we will ascribe 'truth' to (syntactic) formulae, that are interpreted as events, rather than to (semantic) events themselves, but for now that distinction is auxiliary to our main concern.) We also write $\neg E$ to mean $W - E$, and we think of inclusion as implication: if $E \subseteq F$, it means that E *implies* F (in the context of the model). The *event* that E implies F can thus be written $F \cup \neg E$.

In what follows, the properties that we will take to be 'relevant' in possible worlds, are just the strategies chosen by each player. This is entirely without loss of generality: in principle we could include further information, but to keep things simple we stick to just representing the strategies in the model. In order to say that a given model $\mathcal{M}$ is really a model of a particular game $\mathcal{M}$, we will require that for every strategy profile s in the game, there is a state in the model where s is realised.

The information possessed by a player i in a relational model is represented by the relation R_i between states. If sR_it, it means that, *if* the actual state were s, i would consider it possible that the state is t. As we will see, this naturally induces a 'belief operator'.

Definition 1.4. given a game $G = (T_i, \geq_i)_{i \in N}$, a ***relational belief model for*** G is a tuple $(W, (R_i)_{i \in N}, s)$ with W a set, each $R_i \subseteq W \times W$, and $\xi : W \to T$.

If the function ξ is surjective, (i.e. $\forall s \in T \exists u \in W : \xi(u) = s$), then we say that the model is ***full*** for G.

We write $\xi_i(u)$ to mean $(\xi(u))_i$, and we overload notation and lift these functions relations to the power set 2^W of the domain:

$$\xi(E) = \{\xi(u) \mid u \in E\};$$

$$R_i(E) = \{t \in W \mid \exists s \in E : sR_it\}.$$

Sometimes for clarity we use square brackets, e.g. $\xi[E]$ for the lifted forms of these functions. For $u \in W$, we write $R_i(u)$ to mean $R_i(\{u\})$. $R_i(u)$ can be thought of as the 'core' of i's information at the state u, and we'll sometimes call it i's 'information' at u.

Given an event E, we will be interested to define the proposition that *player i believes E*. This is going to be given by an operator $\Box_i$ from events to events, and the operator is just the '*modality*' corresponding to the relation in the model:

$$\Box_i E = \{u \in W \mid R_i(u) \subseteq E\}.$$

So the meaning of $\Box_i$, given the intended interpretation of R_i, is: $\Box_i E$ is true when E is true at every state that player i considers possible.

Definition 1.5. There are a number of restrictions that one sometimes places on properties of beliefs:

T. $\Box_i E \subseteq E$,

D. $\Box_i \emptyset = \emptyset$,

4. $\Box_i E \subseteq \Box_i \Box_i E$,

5. $\neg \Box_i E \subseteq \Box_i \neg \Box_i E$,

mT. $\Box_i (E \cup \neg \Box_i E)$,

mT$^-$. $\Box_i \Box_i E \subseteq \Box_i E$,

where we write $\neg E$ to mean the complement of E in the space, i.e. $\neg E = W - E = \{u \in W \mid u \notin E\}$.

These properties are understood as being implicitly universally quantified (so with a second-order quantifier over events: T for example is to be read 'for all events E, $\Box_i E \subseteq E$). The property T corresponds to i's beliefs being *correct*, i.e. to the *factivity* of $\Box_i$, and so T would be a minimal requirement for us to say that we are talking about 'knowledge' rather than 'belief'. D, which of course is entailed by T, just means that the beliefs of i are always *consistent*. 4 and 5 on the other hand are about the ability of a player to 'introspect': If 4 holds then it means that player i can 'positively introspect', so that if she believes E then she believes[12] that she believes it; and if 5 holds then i can 'negatively introspect': if i doesn't believe E then she believes that she doesn't believe it. mT expresses a kind of 'confidence' in beliefs, saying that a player believes that if she believes something then it is true. It also entails a kind of introspection: that a player is always 'positively' correct about her own beliefs, so that if she believes that she believes something, then she does indeed believe it. mT$^-$ expresses a weaker property than mT: that if player i believes that she believes something, then she does indeed believe it. We might therefore paraphrase mT$^-$ as: 'player i is *correct* about her beliefs.'

[12]In cases of introspection it might be more natural or appropriate to say 'knows', since clearly the relevant belief is true, but we prefer to maintain the same terminology as we use elsewhere in this Chapter.

Definition 1.6. The properties in Definition 1.5 are equivalent[13] to properties of the relations R_i of the underlying relation:

T$^{\text{r}}$. uR_iu,

D$^{\text{r}}$. $R_i(u) \neq \emptyset$,

4$^{\text{r}}$. $R_i(R_i(u)) \subseteq R_i(u)$,

5$^{\text{r}}$. $\{v, w\} \subseteq R_i(u) \Rightarrow w \in R_i(v)$,

mT$^{\text{r}}$. $uR_iv \Rightarrow vR_iv$,

mT$^{\text{-r}}$. $uR_iv \Rightarrow \exists w \in R_i(u) : wR_iv$.

Here the properties are again understood as universally quantified statements, with first-order quantifiers over the variables that stand for states. What this 'equivalence' means is that for example T holds as a universally second-order quantified statement about some model $\mathcal{M}$ if and only if T$^{\text{r}}$ holds as a universally first-order quantified statement about $\mathcal{M}$.

The correspondences for 4,5 and mT make it easy to see that if a player is positively and negatively introspective then she is also confident about her own beliefs. It is even more straightforward to see that T also implies mT. Finally, it is now also easy to see that mT implies mT$^{\text{-}}$.

Fact 1.3. *There are some (well-known) entailments between these properties, for example:*

- *If 4$^{\text{r}}$ and 5$^{\text{r}}$ hold then mT$^{\text{r}}$ holds.*
- *If T$^{\text{r}}$ holds then mT$^{\text{r}}$ holds.*
- *If mT$^{\text{r}}$ holds then mT$^{\text{-r}}$ holds.*
- *If 5$^{\text{r}}$ holds then mT$^{\text{-r}}$ holds.*

Proof. We prove the first on the list, which is the least obvious: Suppose that 4$^{\text{r}}$ and 5$^{\text{r}}$ hold. Then take some $v \in R_i(u)$; by 5$^{\text{r}}$, $u \in R_i(v)$; but then $v \in R_i(R_i(v))$, so by 4$^{\text{r}}$, $v \in R_i(v)$. ■

The introspection property we will introduce in the next Section will be equivalent, on relational models, to mT$^{\text{-}}$.

[13] This equivalence would be 'frame correspondence' in the technical sense of [Benthem, 1976] were we dealing with syntactic versions as in Chapter 2 below.

When a model satisfies T^{r} and 4^{r}, we say that it is 'S4', and when it satisfies in addition 5^{r}, it is called an 'S5' model.[14] In the latter case, each relation defines a partition of the state space, i.e.

$$\{R_i(u) \mid u \in W\}$$

is a partition. Therefore we also call S5 relational models 'partitional models'. These were the models used by Aumann [1976] in his seminal work formalising the notion of common knowledge

We also are interested not just in formalising single-player notions of belief, but in the more truly multi-player notions of *mutual* and *common* belief, and so these are the next items to be formalised.

Definition 1.7. The event that it is ***mutual belief*** of E amongst the players, denoted $\Box E$, is just the event that all players believe E:

$$\Box E = \bigcap_{i \in N} \Box_i E.$$

If a player i is positively and negatively introspective then iterations of the (individual) belief operator i do nothing, i.e. $\Box_i \Box_i E = \Box_i E$. However, the interpersonal notion of mutual belief *does* in general still change with iterations, irrespective of whether or not we impose any of the conditions T, D, 4 or 5: that is, the event that everybody believes that everybody believes that E is not the same as the event that everybody believes that E.

Since we'll be interested in infinite iterations in this Chapter, we define for any arbitrary ordinal $\alpha \in \mathcal{ON}$ the event that there is α-order mutual belief in E.

Definition 1.8. The event that there is α-order mutual belief in E is written $\Box^\alpha E$, and is defined recursively as follows:

$$\begin{array}{ll} \Box^0 E & = \top \\ \Box^1 E & = \Box E \\ \Box^{\beta+1} E & = \Box^\beta E \cap \Box \, \Box^\beta \, E \text{ for } \beta > 0 \\ \Box^\lambda E & = \bigcap_{\alpha<\lambda} \Box^\alpha E \text{ for limit ordinals } \lambda. \end{array}$$

Note that it might conceivably be taken to be objectionable that we here define the $\alpha + 1^{\text{th}}$ level of mutual belief to imply the α^{th} level. In case any defence of this should be needed, let us say a few words. Firstly, it should be clear that in the case when we are modelling *true* belief (or knowledge), taking the intersection makes no difference, since if the $\alpha + 1^{\text{th}}$ level belief is *correct* (i.e. true) then it will entail the previous level, α^{th} level belief, since that is its object. And secondly, the only reasonable definition of the limit case is to take the intersection of all previous levels, and if we defined the

[14] The names 'S4' and 'S5' are from a classification of modal logics that originated with C. I. Lewis; some of the names for axioms are more recent.

lower levels differently then this would be anomalous. Finally, as a step towards a definition of *common belief*, our definition coincides with definitions from the literature, e.g. [Heifetz, 1996, p.111].

Increasing orders of mutual belief are in a certain sense approximations of the next notion that we introduce: common belief.Common belief in E just means arbitrary levels of mutual belief in E. So one natural definition of common belief, a version that we will denote $\Box^\infty$, is the following:

Definition 1.9.

$$\Box^\infty E = \bigcap_{\alpha \in \mathcal{ON}} \Box^\alpha E.$$

However, since clearly $\alpha \geq \beta$ implies that $\Box^\alpha E \subseteq \Box^\beta E$, then for any particular model $\mathcal{M}$ there is some least $\beta_\mathcal{M}$ such that for any event E,

$$\Box^{\beta_\mathcal{M}} E = \Box^{\beta_\mathcal{M}+1} E.$$

This in turn implies that $\Box^{\beta_\mathcal{M}}$ in effect *is* the common belief operator, *in this model*, so that in $\mathcal{M}$, for any event E:

$$\bigcap_{\alpha \in \beta_\mathcal{M}} \Box^\alpha E = \Box^\infty E.$$

This remark about the existence of some $\beta_\mathcal{M}$ will continue to hold even with respect to the larger class of neighbourhood models, since it does not rely on any properties of the belief operator. More remarkably, in the case of relational models we do not need to go beyond the first infinite ordinal ω_0. That is, for relational models $\mathcal{M}$, Fact 1.4 tells us that $\beta_\mathcal{M} \leq \omega_0$.

Fact 1.4. *For any event E and ordinal $\alpha \geq \omega_0$, on relational models we have the following equivalence:*

$$\Box^\alpha E = \Box^{\omega_0} E.$$

We will also define the event $\Box^* E$, that $\Box^*$ is *finitary* common belief.

Definition 1.10.

$$\Box^* E = \bigcap_{m \in \mathbb{N}} \Box^m E.$$

Clearly, by Fact 1.4, on relational models $\Box^* E = \Box^\infty E$: that is, there is no way to distinguish between finitary and absolute common belief on relational models. Fact 1.4 does *not* hold in the more general *neighbourhood models* that we look at in Section 1.4, and indeed $\Box^* E$ and $\Box^\infty E$ will not be the same there. That is why we introduce both Definitions 1.9 and 1.10

There is another characterisation of common true belief, in terms of the existence of a so-called 'evident' event.

Definition 1.11. Call an event 'evident' (for the players N) if it entails that it is believed, i.e. E is ***evident*** when $E \subseteq \Box E$.

Then we have another characterisation of common true belief:

Fact 1.5. *For any event $A \subseteq W$, $u \in \Box^\infty A$ iff there is an evident event E such that $u \in E \subseteq \Box A$.*

This is roughly the form in which Aumann [1976] defined common knowledge formally; this formulation is due to [Monderer and Samet, 1989]. The equivalence relies essentially on the observation that $\Box^\infty E$ is a fixpoint for the $\Box$ operator, and so depends on the *monotonicity* of the $\Box$ operator.[15]

In general the models we have defined allow for the possibility that players are not correct about their own choice of strategy; that is, they might not know what they are doing. (We use the word 'know' here for purely stylistic reasons: we have not forgotten that we prefer to reserve 'knowledge' to mean something stronger than 'true belief' simpliciter.) We might want to rule out this case, but in order to do so we should know what *we* are doing when giving a model of the players' beliefs!

At the moment when they are first presented with the one-shot interaction situation that the game pretends to capture, the players presumably do not know *already* what they will do. Working out what they will do – what strategy choice they will make – should require some reasoning on the part of the players (including, of course, reasoning about what the other players will do). Because we have in mind a deductive interpretation of the one-shot game situation, we should say that in any model of the initial situation, players would have no beliefs concerning what they or the other players will do, or what the beliefs of the other players are. Part of what we will do in Chapter 3 is to look more closely at the deliberative process itself, so considering *intermediate* models representing stages of the process, and at transitions between them. But for our present purposes we generally consider (though none of our results rely essentially on it) that players have all made up their minds about what they will do.

In these situations, we want players to have correct beliefs about what they will do. When the following holds for all players $i \in N$ and all strategies $s_i \in T_i$:

$$\xi^{-1}(s_i) = \Box_i \xi^{-1}(s_i),$$

we say that 'players are correct about their (own) strategies'. The idea of such a model is that it represents the beliefs of the players just before they find out what the other players will do.

It will be convenient for us to define mutual belief in an alternative way, and show that, on relational models, it coincides with the definition already given. Since we must give it a name, let us call it 'co-mutual belief':

[15] Indeed, it is generalised as [Heifetz, 1996, Proposition 2.1] to the case of monotonic neighbourhood models that we will consider in the next Section.

Definition 1.12. The event that an event E is ***α-level co-mutual belief*** is denoted $\boxdot^\alpha E$, and defined recursively as follows:

$$\begin{array}{ll} \boxdot^0 E & = \top \\ \boxdot^1 E & = \Box E \\ \boxdot^{\beta+1} E & = \Box(E \cap \boxdot^\beta E) \text{ for } \beta > 0 \\ \boxdot^\lambda E & = \bigcap_{\alpha<\lambda} \boxdot^\alpha E \text{ for limit ordinals } \lambda. \end{array}$$

We have no argument that co-mutual belief is in itself an entirely natural concept. It is closely related to a notion of 'common knowledge' introduced by Lismont [1994, p. 292]. Note also that *on relational models* it is equivalent to the relatively standard mutual belief as defined above.

Fact 1.6. *Take any relational model and any event E in it. Then $\boxdot^\alpha E = \Box^\alpha E$.*

Proof. Notice first of all that by Fact 1.4, we need only to prove the equivalence $\boxdot^m E = \Box^m E$. But this first term can be written as follows:

$$\underbrace{\Box(E \cap \Box(E \cap \Box(E \cap \ldots \Box E)\ldots))}_{m \text{ times}}.$$

And, in relational models, the following equation holds:

K. $\Box(E \cap F) = \Box E \cap \Box F$.

Therefore our term can be written as:

$$\Box E \cap \Box \underbrace{\Box(E \cap \Box(E \cap \ldots \Box E)\ldots)}_{m-1 \text{ times}};$$

and by repeating this we arrive at

$$\Box E \cap \Box\Box E \cap \Box\Box\Box E \cap \ldots \cap \underbrace{\Box\ldots\Box}_{m \text{ times}} E,$$

which is precisely $\Box^m E$. ■

Notice that we really do need to use the fact that the players' belief operators do respect the equality expressed in K. And indeed in neighbourhood models, where K. does not in general hold, the two definitions do *not* in general coincide. We will only use α-order co-mutual belief in the proof of Theorem 1.3, where this stronger notion than α-order mutual belief is required when we make essentially no assumptions *at all* about the belief operators.

The most important event that we want to define here is *instrumental rationality*. Since we're not interested here in other, more social, aspects of rationality than that generally considered by game theorists, we'll just talk about rationality *tout court*. That fits with contemporary usage by game theorists, for example:

> "*A person's behaviour is **rational** if[f] it is in **his** best interests, given **his** information.*" [Aumann, 2006]

A player's "best interests" are captured by her optimality operator, so that rationality will be parametrised by whatever notion of optimality it involves. So, given that the player's "information" in a relational model is given by the relation R_i, we use Definition 1.13 for rationality.

Definition 1.13. the event that a player i is ***rational*** in a relational model is $\mathbf{r}_i$:

$$\mathbf{r}_i = \{u \in W \mid s_i(u) \in O_i(\xi[R_i(u)])\}.$$

Rationality of a player is the event that the player plays optimality with respect to her information. To repeat using slightly different words: i is rational at u if i's choice of strategy at u is optimal in the context of the restriction defined by i's beliefs at u. (Definition 1.16 below in Section 1.4 defines rationality for neighbourhood models, but that definition will be faithful to this one, in the sense that if we think of a relational model as its equivalent neighbourhood model, then the two definitions coincide.) The event that all players are rational we then write $\mathbf{r}$:

$$\mathbf{r} = \bigcap_{i \in N} \mathbf{r}_i.$$

If each player i has some non-trivial belief about her own strategy, then the definition of rationality only really makes sense when we consider the global versions of operators. This becomes especially clear in the particular case where players correctly believe their own strategies, which as we have suggested is a natural assumption in the scenario being modelled. Suppose that, in some relational model, player i plays s_i and correctly believes that she plays s_i, and does not believe that she plays any other strategy.[16] Then in the restriction defined by her beliefs, the only strategy she plays is s_i. But in the *local* version of any natural optimality operator (including all of the examples we gave in Section 1.1), *player i would always be rational*.

What we really want rationality to mean is that the player plays optimally *among the strategies available to her in the actual game* with respect to her beliefs. And this is precisely what is delivered by considering the global version of any of the optimality notions.

We always therefore assume, for the rest of this Chapter, that the optimality operator under consideration is 'global'.

Definition 1.14. A ***global optimality operator*** for i, O_i, is an optimality operator for i (a function from $2^{T_{-i}}$ to $2TT_i$)with the following property:

$$\text{If } S_{-i} = S'_{-i}, \text{ then } O_i(S) = O_i(S').$$

[16] We have to add that last clause to handle the case of some non-relational neighbourhood models.

There might be further constraints that should be placed on global optimality operators, but this will be sufficient for our purposes.

All the pieces are now in place to state the first Theorem. Theorem 1.1 states that common true belief of rationality entails that players will not play strategies that can be eliminated by iterating the relevant optimality operator.

Theorem 1.1. *In any relational model of a game G, and for any monotonic (global) optimality operator:*

$$\xi[\mathbf{r} \cap \Box^{\infty}\mathbf{r}] \subseteq O^{\infty}.$$

Proof. Take a strategy profile $s \in \xi[\mathbf{r} \cap \Box^{\infty}\mathbf{r}]$; then for some $u \in \mathbf{r} \cap \Box^{\infty}\mathbf{r}$, $\xi(u) = s$. Since $u \in \Box^{\infty}\mathbf{r}$, then by Fact 1.5, there is some $F \ni u$ with $F \subseteq \Box F \cap \Box\mathbf{r}$.

We have the following Lemma.

Lemma 1.1. *$\xi(F \cap \mathbf{r})$ is a post-fixpoint for O, i.e.*

$$\xi(F \cap \mathbf{r}) \subseteq O(\xi(F \cap \mathbf{r})).$$

So by Fact 1.1.2, we have $\xi(F \cap \mathbf{r}) \subseteq O^{\infty}$. Then since $u \in F \cap \mathbf{r}$, we have $s = \xi(u) \in O^{\infty}$ as required.

It suffices then to prove the Lemma. Take any $v \in F \cap \mathbf{r}$ and $i \in N$. Since $v \in \mathbf{r}$, $\xi_i(v) \in O_i(\xi(R_i(v)))$. But $F \subseteq \Box F$, so $R_i(v) \subseteq F$; and $F \subseteq \Box\mathbf{r}$, so $R_i(v) \subseteq \mathbf{r}$; thus $R_i(v) \subseteq F \cap \mathbf{r}$. Therefore by the monotonicity of O_i, we have $\xi_i(v) \subseteq O_i(\xi(F \cap \mathbf{r}))$. ■

The 'converse' direction of Theorem 1.1 states that for any model G, there is a *full* model of G in which wherever any strategy s that survives the iterated elimination of non-optimal strategies is played, there is common true belief of rationality. The Theorem states further that there is such a model in which players are correct about their own strategies. Since we can think of this as the outcome of a process of deliberation and reasoning on the part of the players, it is natural that we ask that players have made up their mind about what they are going to do. (And of course adding this restriction only makes the Theorem stronger.) Since this Theorem also says that the relevant model is a *partitional* model, in which players are positively and negatively introspective, but moreover *always correct* in their 'beliefs', it is also compatible with a 'knowledge interpretation' of $\Box$, so could be read as being about knowledge and common knowledge rather than about belief.

Theorem 1.2. *For any game G, There is a full S5 (partitional) model in which players are correct about their own strategies, and where $\xi(\mathbf{r} \cap \Box^{\infty}\mathbf{r}) = O^{\infty}$.*

Proof. We define a full model: $(T, R_i, id)_{i \in N}$ by setting:

$$R_i(s) = \begin{cases} \{t \in T \mid s_i = t_i \ \& \ t_{-i} \in (O^{\infty})_{-i}\} & \text{if } s \in O^{\infty} \\ \{t \in T \mid s_i = t_i \ \& \ t_{-i} \in T_{-i} - (O^{\infty})_{-i}\} & \text{otherwise.} \end{cases}$$

Note here that in this model the worlds (states) are strategy profiles, and ξ is the identity function i.e. $\xi(x) = x$. We must show therefore that in this model, $\mathbf{r} \cap \Box^{\infty}\mathbf{r} = O^{\infty}$. The $\subseteq$ inclusion is Theorem 1.1, so only the $\supseteq$ direction remains:

First, take any $s \in O^{\infty}$; we will show that $s \in \mathbf{r}$: For any player i, by construction $(R_i(s)))_{-i} = (O^{\infty})_{-i}$, so we certainly have $s \in O_i(R_i(s))$, since the operator is global (cf. Definition 1.14).

To show that $s \in \Box^{\infty}\mathbf{r}$, it will suffice, given Fact 1.4, to show that for any m-length sequence $w \in N^m$ of players,

$$R_{w(1)}(R_{w(2)}(\ldots R_{w(m)}(s)\ldots)) \subseteq \mathbf{r}. \tag{1.1}$$

And since, as we have just seen, $O^{\infty} \subseteq \mathbf{r}$, then it will suffice, in order to show 1.1, to show that:

$$R_{w(1)}(R_{w(2)}(\ldots R_{w(m)}(s)\ldots)) \subseteq O^{\infty}.$$

In order to see this last inclusion, notice that for any player i, if $E \subseteq O^{\infty}$ then $R_i(E) \subseteq O^{\infty}$. ■

The weaker form in which Theorem 1.2 is usually stated is an immediate corollary of it that we state just in case it can make some connections (with the literature and with the contents of Section 1.4) clearer.

Corollary 1.1. *For any strategy s in the game G that does survive the iterated elimination of non-optimal strategies, there is a* full *model of G in which at some state s is played* and *there is common true belief of rationality.*

1.4 Transfinite mutual belief in rationality

Tan and Werlang [1988] proved their theorem about mutual belief and rationality in terms of (finite) stages, meaning that there is an 'equivalence' between the level of mutual belief of rationality and the number of rounds of iteration of elimination of non-optimal strategies. They in effect proved the following:

Theorem. *The following are equivalent for $m \in \mathbb{N}$:*

1. *The strategy profile s in the game G survives m rounds of iteration of strictly dominated strategies;*
2. *There is a model of G with a state $u \in \mathbf{r} \cap \Box^{m}\mathbf{r}$ such that $\xi(u) = s$, where $\mathbf{r}$ means the players avoid strategies they believe to be strictly dominated.*

In this Section we strengthen that result. Firstly, as before, we consider other kinds of (monotonic) optimality than just not playing a strictly dominated strategy. Secondly, we consider the more general transfinite case. That is going to mean considering models that are not relational: so-called 'neighbourhood models' and a special case: 'topological models'. Finally, we shift a quantifier, so prove a stronger result. That is, where

the *1* $\Rightarrow$ *2* direction of the above Theorem states that 'for any $m \in \mathbb{N}$ and strategy profile $s \in O^m$, there is a model such that ...', we will prove a result of the form 'there is a model such that, for any $\alpha \in \mathcal{ON}$ ...'.

The other direction of the implication says that α-level mutual belief in rationality entails not playing strategies that are eliminated by α rounds of elimination of non-optimal strategies. We prove this result in two forms, since we cannot prove the result as it stands with respect to the most general class of neighbourhood models. The first form, Theorem 1.3 replaces 'mutual belief' with 'co-mutual belief'. Therefore, given that on relational models co-mutual belief *is* mutual belief, Theorem 1.1 above is a corollary of 1.3, but we still gave a separate proof of Theorem 1.1 using fixpoints, to illustrate its simplicity. The second form, Theorem 1.4, adds a specific kind of introspection condition to the neighbourhood models. (That condition is satisfied on many models, including topological models.)

In partial motivation of this Section, we will first show that there are games requiring transfinite rounds of elimination of non-optimal strategies. Then we introduce neighbourhood models, show the connection with relational models, define rationality on neighbourhood models, and state Theorem 1.3, that looks at the implications of transfinite levels of co-mutual belief of rationality. After that we introduce topological models, and state some well-known equivalences between topological models and neighbourhood models, and between some classes of topological models and relational models. We then give Theorem 1.4, that looks at the implications of levels of mutual belief given a particular introspection condition.

This leads us to Theorem 1.5, the 'converse' to Theorem 1.4, in which we show that there is a model in which α-level mutual belief of rationality is strictly *equivalent* to α rounds of elimination of non-optimal strategies. To end this Section, and the first Chapter, we look at the case of S5 models, and remark that our strong formulation of Theorem 1.5 does not hold with respect to S5 models.

For some games we really need to complete an infinite number of rounds of elimination of non-optimal strategies. Consider for example, in the pick-the-highest-number game, in which two players a and b, must pick a number in ignorance of what the other has picked. The preferences over outcomes are that each player i strictly prefers picking a (strictly) higher number than the other, and is indifferent between other factors. Then 0 is dominated for both players, since they strictly prefer 1 in all situations. But then they strictly prefer 2, and so one... Clearly there is no number that is not strictly dominated, so that infinite rounds of elimination yield the empty restriction $\emptyset$. That means by Theorem 1.1 that if they commonly believe each other to be rational then there is nothing they can rationally play. (Of course, at any *finite* stage of iteratively eliminating, there are still strategies left in the resulting restriction.)

We will show that there are also games which require infinite rounds of elimination of non-optimal strategies. In particular, we show that for *any* ordinal α, there is a game that requires α rounds of elimination of strictly dominated strategies in order to reach

the outcome, that in addition is non-empty.[17]

Fact 1.7. *Given an arbitrary ordinal β, we can construct a game G_β which has outcome ordinal β, and which is such that O^∞_G is nonempty, where the optimality notion is that of strict dominance.*

Proof. Let G_β be the two-player strategic game with $\{0, 1, \ldots, \beta, \beta+1\}$ as strategies, and preferences determined by the following payoff functions:

$$\pi_i(s) = \begin{cases} 0 & \text{if } s_{-i} \geq s_i \neq \beta + 1 \\ 1 & \text{otherwise} \end{cases}$$

This is a simplification and generalisation of [Chen *et al.*, 2007, Example 1]. The reader can check that in this game it takes precisely β rounds of iteration in order to reach the fixpoint outcome; i.e. that $\alpha_{G_\beta} = \beta$. (On this game the local and global operators coincide.) The idea is that in the first round, 0 is strictly dominated by $\beta + 1$ (in fact it happens to be weakly dominated by all the other strategies), but 1 is 'safe' even in the presence of the strictly dominating $\beta+1$ because the other player is allowed to choose 0; $\beta + 1$ is rationalisable on the grounds that the opponent might play 0. However, once 0 has disappeared, 1 *is* then strictly dominated, but 2 is not (yet) because the opponent might play 1, and so on ... ■

Furthermore, this is no 'special feature' of strict dominance. Indeed, the reader interested in playing with these things can also check that Fact 1.7 holds for the cases of weak dominance and best response.

As we have already indicated, over relational models finitary common belief $\Box^*$ coincides with absolute common belief $\Box^\infty$. In the general case then, for arbitrary (possible transfinite) ordinals α, it is not possible to give a model in which for all α, $1 + \alpha$ rounds of elimination entail rationality and mutual belief in rationality. What is more, even the weaker $\forall\exists$ form of the implication that we might like fails.

Indeed, as soon as (strictly) more than ω_0 steps are required for the game's outcome to be reached, there is no way to give the required model. To see this, suppose that for some game G and (collective) optimality operator O, more than ω_0 rounds of elimination are required to reach the outcome. Then there is some $s \in O^{\omega_0} - O^{\omega_0+1}$. Now suppose that Tan and Werlang's Theorem above held for arbitrary ordinals with respect to *relational* models. Then by the 1⇒2 direction of the Theorem, there would be a *relational* model $\mathcal{M}$ with some state u in it such that $\xi(u) = s$ and $u \in \mathbf{r} \cap \Box^{\omega_0}\mathbf{r}$. By Fact 1.4, we know that $\Box^{\omega_0}\mathbf{r} \subseteq \Box^{\omega_0+1}\mathbf{r}$, so $u \in \mathbf{r} \cap \Box^{\omega_0+1}\mathbf{r}$, in which case by the 2⇒1 direction, would have $u \in O^{\omega_0+1}$, which contradicts our initial assumption. Therefore Tan and Werlang's Theorem cannot be extended to arbitrary ordinals with respect to relational models.

Recall that we said that a *neighbourhood model* contains a *list* of pieces of information possessed by a player.

[17] Actually similar examples can also be constructed for all of the other optimality operators we considered in Section 1.1; we consider the case of non-strict dominance as an illustration.

Definition 1.15. a ***neighbourhood model*** for the game G is a structure $(W, (\mathcal{N}_i)_{i\in N}, \xi)$, with W and ξ are as in Definition 1.4 of relational models, and $\mathcal{N}_i$ a function associating with each state $u \in W$ a set of events; so $\mathcal{N}_i : W \to 2^{2^W}$, with the only condition that $W \in \mathcal{N}_i(u)$ for every $u \in W$.

How does one define belief in neighbourhood models? We simply say that a player believes the event E at u just if $E \in \mathcal{N}_i(u)$:

$$\begin{array}{rcl} \Box_i E & = & \{u \in W \mid E \in \mathcal{N}_i(u)\} \\ & \text{i.e.} & \\ u \in \Box_i E & \text{iff} & E \in \mathcal{N}_i(u). \end{array}$$

Notice that the restriction that we placed on the neighbourhoods, that they must include W, then means that $\Box_i W = W$ (as is also the case for relational models). the levels of mutual belief, and the two kinds of common belief, are now defined as before, given this new definition. So α-order mutual belief of E is defined inductively, just as in Definition 1.8, as follows:

$$\begin{array}{ll} \Box^0 E & = \top \\ \Box^1 E & = \Box E \\ \Box^{\beta+1} E & = \Box^\beta E \cap \Box\, \Box^\beta\, E \text{ for } \beta > 0 \\ \Box^\lambda E & = \bigcap_{\alpha<\lambda} \Box^\alpha E \text{ for limit ordinals } \lambda. \end{array}$$

Common belief of E, which we will still write $\Box^\infty E$, is defined as in Definition 1.9:

$$\Box^\infty E = \bigcap_{\alpha \in \mathcal{ON}} \Box^\alpha E.$$

And finitary common belief of E is the same as in Definition 1.10:

$$\Box^* E = \bigcap_{m \in \mathbb{N}} \Box^m E.$$

These two definitions of common belief ('finitary' and 'absolute') now do *not* coincide, given the new definition of the underlying concept of belief that we have for neighbourhood models. This fact will be a corollary of Theorem 1.5.

The definition that we had for rationality needs also to be reworked. We now think of the events in i's neighbourhood of a point as i's information at that point. We will see later the standard results connecting neighbourhood models with relational models, but for now notice just that there are fewer constraints on how the player has put her information together. Still, we can define rationality on neighbourhood models: we just say that a player is rational if she acts in her best interests according to *all* of her information.

Definition 1.16. The event that player i is ***rational*** in a neighbourhood model is $\mathbf{r}_i$:

$$\mathbf{r}_i = \{u \in W \mid \forall A \in \mathcal{N}_i(u),\, \xi_i(u) \in O_i(\xi(A))\},$$

and again the event that all players are rational $\mathbf{r}$ is just $\bigcap_{i\in N} \mathbf{r}_i$.

It is immediate from inspecting the translation we give below in Fact 1.8 from relational models to neighbourhood models that this captures the same notion as in relational models, so Definition 1.16 is properly speaking a generalisation of Definition 1.13 of rationality on relational models. To our knowledge, this is the first formulation of rationality for neighbourhood models.

We will be interested, to a certain extent in this Chapter, but also in the next, in *monotonic* neighbourhood models. For players' beliefs, this would mean that if a player believes an event A and some other event E is entailed by A (i.e. $A \subseteq E$), then the player believes E as well. A ***monotonic neighbourhood model*** is a neighbourhood model in which each of the neighbourhood functions satisfies the monotonicity property M:

M. If $A \in \mathcal{N}_i(u)$ and $A \subseteq E$ then $E \in \mathcal{N}_i(u)$.

Monotonic neighbourhood models are studied from a logical perspective in [Hansen, 2003], where many model-theoretic results are established. They are also studied, with special attention being paid to *common belief* in [Heifetz, 1996]. Although monotonicity is a natural enough requirement, and is often technically useful (for example it will be required for a number of results in Chapter 3), we do *not* require the neighbourhood models we consider, for Theorems 1.3 or 1.4, to be monotonic.

Since neighbourhood models really are just lists of beliefs, it will be necessary, if we are to prove much about them, to introduce a restriction on how beliefs behave. Specifically, we will want to impose the following condition:

mT$^+$ $\Box_i(\Box_i E \cap F) \subseteq \Box_i(E \cap F)$.

This condition mT$^+$ is equivalent on relational models to the condition mT$^-$, that players are correct about their own beliefs in the sense that if a player believes she believes something, then she does indeed believe it. So by Fact 1.3 5 entails mT$^+$ on relational models. However, on monotonic neighbourhood models mT$^+$ is a strictly stronger condition than both mT$^-$ and 5. (On neighbourhood models mT$^+$ is still entailed by the condition T, that players' beliefs are correct.)

We do not claim to have a natural way of reading the condition mT$^+$, but it certainly is difficult to find an argument against it, given that it says that 'if you believe that: F and you believe that E, then you believe E and F.' This is the condition that will be required to prove Theorem 1.4, which will state that rationality plus α-level mutual belief of rationality entail not playing strategies that are eliminated by $1 + \alpha$ rounds of elimination of non-optimal strategies.

Fact 1.8 says that we can think of relational models as neighbourhood models.

Fact 1.8. *For every relational model there is a neighbourhood model that is equivalent to it, in the sense that it has the same state space W and strategy function ξ, and the belief operators $\Box^\alpha$ are the same.*

Proof. Given any relational model $(W, R_i, \xi)_{i \in N}$, the monotonic neighbourhood model that is equivalent to it, in the sense just described, is simply $(W, \mathcal{N}_i, \xi)_{i \in N}$, where

$$\mathcal{N}_i(u) = \{E \subseteq W \mid R_i(u) \subseteq E\}.$$

A quick inspection of the definitions suffices to establish the equivalence between this model and the previous one. ■

Given the existence of an equivalent monotonic neighbourhood model for every relational model, we will sometimes (especially in the next Chapter) talk interchangeably about a relational model and its neighbourhood version, so that we sometimes speak of relational models as *being* neighbourhood models.

The reverse translation clearly does not in general exist: there are monotonic neighbourhood models for which there exists no equivalent relational model. The smallest example is the 1-player model $(W, \mathcal{N}_a, \xi)$ given by:

$$\begin{array}{l}(W = \{u, v\},\\ \mathcal{N}_a(u) = \{\{u\}, \{v\}, \{u, v\}\}\\ \mathcal{N}_a(v) = 2^W,\\ \xi).\end{array}$$

Here we have $u \in \Box_a\{u\}$ and $u \in \Box_a\{v\}$. Suppose that there were some relational model (W, R_a, ξ) equivalent to the given model. Then we would have to have $R_a(u) \subseteq \{u\}$ and $R_a(u) \subseteq \{v\}$, meaning that $R_a(u) = \emptyset$. But then we'd have $u \in \Box_a\emptyset$, which does *not* hold in the original model.

Therefore, a restriction must be placed on neighbourhood models if they are to be equivalent to relational models. That restriction is that the neighbourhoods must be monotonic and each *contains its core*.

Definition 1.17. We say that the model $(W, \mathcal{N}_i, \xi)_{i \in N}$ ***contains its core*** iff

$$\forall u \in W, \bigcap \mathcal{N}_i(u) \in \mathcal{N}_i(u).$$

For any monotonic neighbourhood model that contains its core, we can define a relational model that is equivalent to it:

Definition 1.18. Let $(W, \mathcal{N}_i, \xi)$ be a monotonic intersection-closed neighbourhood model (i.e. that contains its core). Then we define the model $(W, R_i, \xi)_{i \in N}$ that is (as is straightforward to see) equivalent to it as follows:

$$R_i(u) = \bigcap \mathcal{N}_i(u).$$

This formal property of 'containing the core' corresponds precisely to the more intuitive notion of 'putting together one's information'. We say that in relational models, each player has put together her pieces of information, and by this we mean that if, at

u, a player believes E and believes F – so $u \in \Box_i E \cap \Box_i F$ – then the player believes $E \cap F$, which can be read 'E and F': $u \in \Box_i(E \cap F)$. In fact relational models assume more: that players are able to put together *infinitely many* pieces of information, so that if they believe all of the propositions

$$\{E_\beta\}_{\beta \in \alpha},$$

then they will also believe the proposition

$$\bigcap_{\beta \in \alpha} E_\beta.$$

In neighbourhood models it is *not* in general the case that players put together even only *finitely* many pieces of information. That is, the following equality, that, as we saw, held for relational models[18], does not hold in general on neighbourhood models.

K. $(\Box_i E \cap \Box_i F) = \Box_i(E \cap F)$.

The first result that we establish in this Section concerns the notion of co-mutual belief (Definition 1.12). Just like Theorem 1.4 that we present immediately afterwards, it 'zooms in' on the individual stages of elimination, so that rather than just saying, as did Theorem 1.1, that the limit case of *common* belief of rationality entails the *outcome* of eliminating non-optimal strategies, we now look at intermediate levels of (just for this Theorem) co-mutual belief, and associate them with corresponding numbers of rounds of elimination of non-optimal strategies.

Theorem 1.3. *In any neighbourhood model* $(W, \mathcal{N}_i, \xi)_{i \in N}$ *of the game* $(T, \leq_i)$*, for any* $\alpha \in \mathcal{ON}$*,* $\xi(\mathbf{r} \cap \boxdot^\alpha \mathbf{r}) \subseteq O^{1+\alpha}$.

Proof. We prove this directly by transfinite induction on α.

0: We must show that $\xi(\mathbf{r}) \subseteq O^1$. Take any $s \in \xi(\mathbf{r})$. Then there is some $u \in \mathbf{r}$ such that $\xi(u) = s$. For each i, $u \in \mathbf{r}_i$, so by definition $\forall U \in \mathcal{N}_i(u), \xi_i(u) \in O_i(\xi[U])$. But because $W \in \mathcal{N}_i(u)$ and O_i is monotonic, we have $\xi_i(u) \in O_i(T)$. Repeating this for all players $i \in N$ yields $\xi(u) \in O(T) = O^1$.

I: By the inductive hypothesis, $\xi(\mathbf{r} \cap \boxdot^\beta \mathbf{r}) \subseteq O^{1+\beta}$. Then take any $s \in \xi(\mathbf{r} \cap \boxdot^{\beta+1}\mathbf{r})$; there is u with $\xi(u) = s$ such that $u \in \mathbf{r} \cap \Box(\mathbf{r} \cap \boxdot^\beta \mathbf{r})$.

Then since $u \in \mathbf{r}$, we have: $\forall U \in \mathcal{N}_i(u)$, $\xi_i(u) \in O_i(\xi[U])$, and since $u \in \Box(\mathbf{r} \cap \boxdot^\beta \mathbf{r})$, we have $\mathbf{r} \cap \boxdot^\beta \mathbf{r} \in \mathcal{N}_i(u)$. So we have $\xi_i(u) \in O_i(\xi[\mathbf{r} \cap \boxdot^\beta \mathbf{r}])$, so by the inductive hypothesis and the monotonicity of O_i, we have $\xi_i(u) \in O_i(O^{1+\beta})$, as required.

[18] Indeed, it is named in honour of Saul Kripke for his work on logical completeness involving relational models.

Λ: The inductive hypothesis states that $\forall \beta < \lambda$, $\xi[\mathbf{r} \cap \boxdot^{\beta}\mathbf{r}] \subseteq O^{1+\beta}$. Then we immediately have the following inclusions:

$$\xi[\mathbf{r} \cap \boxdot^{\lambda}\mathbf{r}] = \xi[\mathbf{r} \cap \bigcap_{\beta<\lambda} \boxdot^{\beta}\mathbf{r}] = \bigcap_{\beta<\lambda} \xi[\mathbf{r} \cap \boxdot^{\beta}\mathbf{r}] \overset{\text{I.H.}}{\subseteq} \bigcap_{\beta<\lambda} O^{1+\beta} = O^{\lambda}.$$

■

Since, as we observed when defining it, co-mutual belief is *equivalent* to mutual belief on relational models, we have the following immediate corollary of Theorem 1.3.

Corollary 1.2. *On* relational *models of G, for any $\alpha \in \mathcal{ON}$, $\xi(\mathbf{r} \cap \Box^{\alpha}\mathbf{r}) \subseteq O^{1+\alpha}$.*

The case of co-mutual belief is perhaps not very instructive: although it coincides with belief on relational models, and it might well have some intuitive appeal on neighbourhood models, this remains unclear. Therefore we now give another formulation of Theorem 1.3, this time in terms of mutual belief itself. Now however we will require that the model satisfy the introspection-like property mT$^+$ given above.

Theorem 1.4. *In any neighbourhood model of G that satisfies mT$^+$, $\xi(\mathbf{r} \cap \Box^{\alpha}\mathbf{r}) \subseteq O^{1+\alpha}$.*

Proof. We again establish the claim itself by transfinite induction on α. The start and limit cases are essentially the same as in the proof of Theorem 1.3, so we give just the successor step:

I. The inductive hypothesis tells us that

$$\xi(\mathbf{r} \cap \Box^{\beta}\mathbf{r}) \subseteq O^{1+\beta},$$

and we want to show that, for any player $i \in N$:

$$\xi(\mathbf{r} \cap \Box^{\beta+1}\mathbf{r}) \subseteq O_i(O^{1+\beta}).$$

Take any $s \in \xi(\mathbf{r} \cap \Box^{\beta+1}\mathbf{r})$; we have u with $\xi(u) = s$ such that $u \in \mathbf{r} \cap \Box^{\beta+1}\mathbf{r}$.

Since $u \in \mathbf{r}$, we know that $\forall A \in \mathcal{N}_i(u)$, $\xi_i(u) \in O_i(\xi[A])$. Therefore, by the inductive hypothesis and the monotonicity of O_i, if we can show that $\mathbf{r} \cap \Box^{\beta}\mathbf{r} \in \mathcal{N}_i(u)$, then we are done. (Because then we would have $\xi_i(u) \in O_i(\xi[\mathbf{r} \cap \Box^{\beta}\mathbf{r}])$.)

So we will use the fact that $u \in \Box^{\beta+1}\mathbf{r}$, in order to show that $\mathbf{r} \cap \Box^{\beta}\mathbf{r} \in \mathcal{N}_i(u)$. We distinguish two cases:

- If $\beta = 0$ then $\mathbf{r} \cap \Box^{\beta}\mathbf{r} = \mathbf{r}$, so we need only to show that $\mathbf{r} \in \mathcal{N}_i(u)$:

$$u \in \Box^{\beta+1}\mathbf{r} = \Box\mathbf{r} = \{v \in W \mid \mathbf{r} \in \mathcal{N}_i(v)\}.$$

- Otherwise, $\beta > 0$. In this case, by Definition 1.8, $\Box^{\beta}\mathbf{r} \subseteq \Box_i\mathbf{r}$. That is: $\Box^{\beta}\mathbf{r} = \Box_i\mathbf{r} \cap \Box^{\beta}\mathbf{r}$. The property mT$^+$ then applies, so that $\mathbf{r} \cap \Box^{\beta}\mathbf{r} \in \mathcal{N}_i(u)$.

■

We have already mentioned *topological models*, and now is the time to introduce them formally. We will give all the definitions needed to understand the material here, but refer to [Munkres, 1999] for an introduction to the subject of topology.

Definition 1.19. A ***topology*** over Ω is a set τ of subsets of Ω that has the following properties:

- for any $X \subseteq \tau$, $\bigcup X \in \tau$;
- for any *finite* $X \subseteq \tau$, $\bigcap X \in \tau$.

The elements of τ are called the ***open*** sets, and if an event $E \subseteq W$ is open then we say its complement $W - E$ is ***closed***. The ***interior*** of an event E, written $int(E)$ is *the largest* open set contained in E, i.e. the union of all of the open sets contained in E, which we can write $\bigcup(2^E \cap \tau)$. The ***closure*** of E is the smallest closed set containing E.

In the multi-player topological semantics that we present, each player i is assigned a topology τ_i over Ω.

Definition 1.20. A ***topological model*** for $(T_i, \geq_i)_{i \in N}$ is thus a tuple

$$(\Omega, \tau_i, \xi)_{i \in N},$$

where $\xi : \Omega \to T$, and each τ_i a topology over Ω.

For $u \in \Omega$, we will write $\tau_i(u)$ to mean the set of i-open sets with u in them, i.e. $\tau_i(u) = \{U \in \tau_i \mid u \in U\}$.

$\Box_i$ is the interior operator, i.e. $\Box_i E = int_i(E)$. Relational models for modal logic were invented in the late 50s, but the topological semantics have been studied considerably before that [McKinsey and Tarski, 1944]. The interior operator is an $S4$ modality, i.e. belief is *factive* (so some might prefer to call it 'knowledge'); players are positively introspective; and each player believes the finitary implications of her beliefs. Common belief (or common knowledge) is studied on topological models in [Benthem and Sarenac, 2004].

We use all the definitions of mutual and common belief as for relational models (so see Section 1.3), except that $\Box_i$ is now defined as the interior operator with respect to the topology τ_i. In topological models, as in monotonic neighbourhood models, we will see that we *can* distinguish between finitary absolute common belief, i.e. $\Box^*$ and $\Box^\infty$ also do not coincide on topological models. We will establish this by proving Theorem 1.5, the converse direction for Theorem 1.4, with respect to topological models.

Topological models can also be thought of as neighbourhood models in which the neighbourhood $\mathcal{N}_i(u)$ of any point u is the monotonic closure of the set of open sets which have u in them, with respect to some topology τ_i. That is, given the topological

model $(W, \tau_i, \xi)_{i \in N}$, we can define a neighbourhood model that is equivalent to it as $(W, \mathcal{N}_i, \xi)_{i \in N}$, where:

$$\mathcal{N}_i(u) = \{E \subseteq W \mid \exists A \in \tau_i(u) : \ A \subseteq E\}.$$

To see the equivalence between this neighbourhood model and the topological model, notice that we have the following definition of the $\Box_i$ operators in the neighbourhood model:

$$\Box_i E = \{u \in \Omega \mid \exists U \in \tau_i : u \in U \subseteq E\}.$$

This is, as we wanted, equivalent to the interior operator. Therefore, just as we talk interchangeably about a relational model and its neighbourhood version, we will also sometimes think of topological models as monotonic neighbourhood models.

Not all topological models are (equivalent to) a relational model, but some are. More precisely, those topological models that are ***Alexandroff***, i.e. in which every point has a unique smallest open set around it, are equivalent to a relational model. Equivalently, Alexandroff topologies are those in which the set of open sets is also closed for arbitrary (rather than just finite) intersections. That is, thinking of topological models as monotonic neighbourhood models, Alexandroff models are those that contain their core (Definition 1.17).

That is because the same property will clearly be carried over the topological model thought of as a neighbourhood model, in which case the translation given above in Definition 1.18 from intersection-closed (i.e. core-containing) monotonic neighbourhood models to relational models can be used to provide an equivalent relational model.

Given a topological model, we can define the *Alexandroff supplementation* of it as the smallest topological model that extends it and in which the topologies are Alexandroff.

Definition 1.21. Given some topological model $(W, \tau_i, \xi)_{i \in N}$, its ***Alexandroff supplementation*** is the topological model $(W, \tau_i', \xi)_{i \in N}$, where:

$$\tau_i' = \{E \subseteq W \mid \exists X \subseteq \tau_i : \ E = \bigcap X\}.$$

In the Alexandroff supplementation, each player now puts together all her information. That is, when thought of as a neighbourhood model, the Alexandroff supplementation contains its core, and so is equivalent to a relational model. In general in topological models, players *do* put their information together finitarily, so that, unlike in neighbourhood models, the equation K does hold. However, what fails is of course its *infinitary* version,

K^{∞}. $\bigcap_{\beta \in \alpha} \Box_i E_\beta = \Box_i (\bigcap_{\beta \in \alpha} E_\beta)$,

and it is precisely this fact that we will exploit now.

Having defined topological models, we are in a position to be able to state our theorem involving them, which is a strong 'converse' to Theorems 1.3 and 1.4.

Theorem 1.5. *For any game G, there is a full topological model of G in which players are correct about their strategies, and where for any ordinal α, we have $\xi(\mathbf{r} \cap \Box^\alpha \mathbf{r}) = O_G^{1+\alpha}$.*

Proof. Let $G = (T_i, \geq_i)_{i \in N}$. We will define a topological model $\mathcal{M}_G$ of G in which $\xi[\mathbf{r} \cap \Box^\alpha \mathbf{r}] = O_G^{1+\alpha}$.

Given a strategy s_i belonging to player i, we write $dp_i(s_i)$ (the 'depth' of s_i) to mean, roughly speaking, the number of rounds it takes before s_i will be eliminated. More formally:

$$dp_i(s_i) = \begin{cases} \alpha_G^O + 1 & \text{if } s_i \in O_i(O_G^\infty) \\ max\{\alpha < \alpha_G^O \mid s_i \in [O^\alpha(T)]_i\} & \text{otherwise.} \end{cases}$$

What this means is that those strategies that are eliminated in the first are each assigned a depth of 0; those eliminated in the second round get a depth of 1, and so on; and those that are never eliminated are assigned a depth of α_G^O. (Recall that α_G^O is the *outcome ordinal* of the game for O, so that if a strategy is not eliminated by that round then it never will be.)

This depth function has the following key property:

Lemma 1.2. *If $dp_i(s_i) > \delta$, then $s_i \in O_i(\{s_i\} \times [O^\delta(T)]_{-i})$.*

Proof. By the definition of dp_i, we have $s_i \in O_i(O^\delta(T))$. Then we use the globality of O_i: since

$$[O^\delta(T)]_{-i} = [\{s_i\} \times [O^\delta(T)]_{-i}]_{-i},$$

then we have (cf. Definition 1.14)

$$O_i([O^\delta(T)) = O_i(\{s_i\} \times [O^\delta(T)]_{-i}).$$

■

We will use this depth function dp_i to define the basis of the topology. The domain of the model is the set of strategy profiles T; the function ξ will be identity; and the topology we define by the following basis for each player i:

$$\begin{aligned} \mathbb{B}_i \;=\; & \Big\{ \{s_i\} \times [O^\beta(T)]_{-i} \Big| s_i \in T_i,\, \beta < dp_i(s_i) \Big\} \\ & \bigcup \; \Big\{ \{s_i\} \times T_{-i} \Big| s_i \in T_i - O_i(T) \Big\}. \end{aligned}$$

τ_i is then generated by taking arbitrary unions from $\mathbb{B}_i$.

It can be useful to see a picture of this model, so we give in Figure 1.6 a depiction of the basis $\mathbb{B}_i$ for the row player i in a two-player game. There we assume, just for the purposes of illustration, that i has one strategy eliminated in each round until the end.

It can be verified that $\mathbb{B}_i$ is indeed a basis. Thus that it induces, by taking arbitrary unions, a topology τ_i, meaning that $(T, \tau_i, id)_{i \in N}$ is indeed a topological model of G.

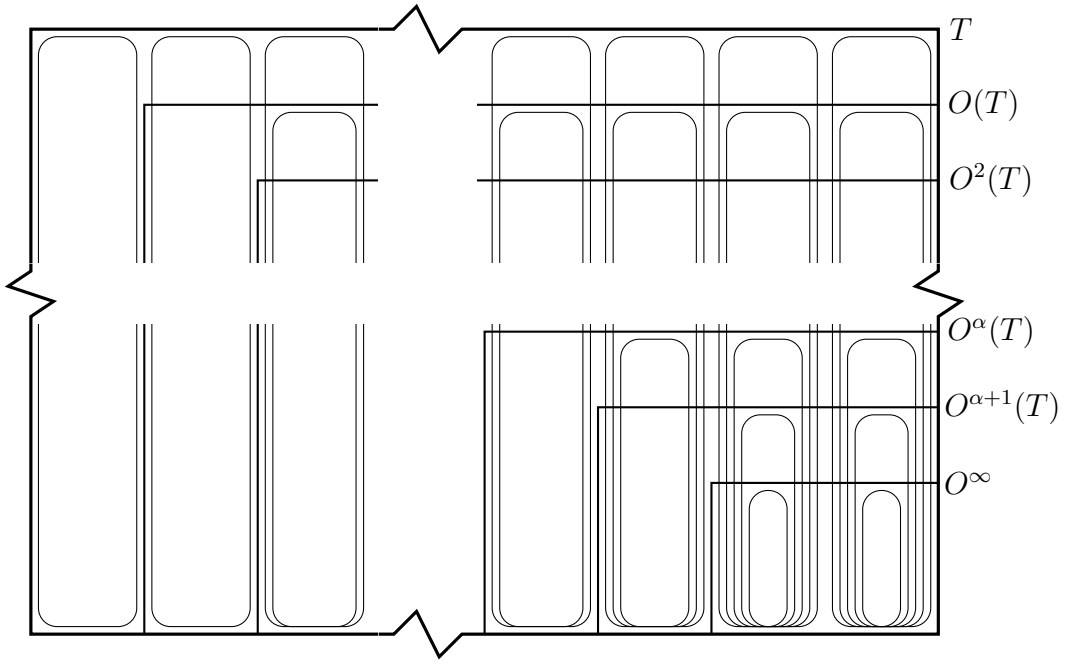


Figure 1.6: An example of the model used to prove Theorem 1.5. We show only the topology for one player, who is choosing which column in the space the outcome will be in.

Then we will prove that in this model $\mathbf{r} \cap \Box^\alpha \mathbf{r} = O^{1+\alpha}$. The $\subseteq$ direction we have by Theorem 1.4, given that (a) since topological models satisfy T, they also satisfy mT$^+$; and (b) in this model ξ is the identity function. The other direction ($O^{1+\alpha} \subseteq \mathbf{r} \cap \Box^\alpha \mathbf{r}$) is by induction:

0: Take $s \in O(T)$ and $i \in N$. We must show that $s \in \mathbf{r}_i$. Take any $A \in \tau_i(s)$. We know that $s_i \notin T_i - O_i(T)$, since $s \in O(T)$. So by definition of the model, there is some $\delta < dp_i(s_i)$ such that $\{s_i\} \times [O^\delta(T)]_{-i} \subseteq A$. Since by the Lemma $s_i \in O_i(\{s_i\} \times [O^\delta(T)]_{-i})$, then by monotonicity of O_i, we have $s_i \in O_i(A)$ as required.

I: Take $s \in O^{1+\beta+1}$. For each player $i \in N$, we have by construction $\{s_i\} \times O_{-i}^{1+\beta} \in \tau_i(s)$. Furthermore, clearly $\{s_i\} \times O_{-i}^{1+\beta} \subseteq O^{1+\beta}$, and by the inductive hypothesis $O^{1+\beta} = \Box^\beta \mathbf{r}$. Therefore $s \in \Box_i \Box^\beta \mathbf{r}$. Repeating this for each player $i \in N$, we have $s \in \Box \Box^\beta \mathbf{r}$. But because $O^{1+\beta+1}(T) \subseteq O^{1+\beta}(T)$, then we have $s \in O^{1+\beta}(T)$ and so the inductive hypothesis $s \in \Box^\beta \mathbf{r}$. Therefore $s \in \Box^\beta \mathbf{r} \cap \Box \Box^\beta \mathbf{r} = \Box^{\beta+1} \mathbf{r}$.

Λ: Immediate from the inductive hypothesis:

$$O^{1+\lambda} = \bigcap_{\beta<\lambda} O^{1+\beta} \overset{\text{I.H.}}{=} \bigcap_{\beta<\lambda} \Box^\beta \mathbf{r} = \Box^\lambda \mathbf{r}.$$

■

Notice that in the finitary case (or in fact when the game's outcome is reached after ω_0 steps), each player's topology is finite and therefore Alexandroff, and so the model defined is relational. Therefore the result *does* hold for finite ordinals α with respect to relational models.

It is worth briefly seeing why a relational version of the model given in Theorem 1.5 does *not*, in the general infinitary case, have the desired property. If we were to take the Alexandroff supplementation of that model, applying Definition 1.21, then the reason the result fails to hold is that we would lose the rationality of the players at limit stages: Take some state u at which player i's strategy is eliminated at the λ^{th} stage, for some limit ordinal λ. Then there will be a smallest neighbourhood $U \in \tau_i(u)$, in which all players play only strategies that are eliminated by λ rounds of elimination. But in that case, i is not rational at u.

The reason i was rational in the (non-Alexandroff) topological case was that she had failed to put together all of her information: At u she has, for any $\beta < \lambda$, and any strategy profile $s \in O^\beta$, information that tells her that s_{-i} will not be played. Yet still she does not put together all of these pieces of information in order to conclude that the other players will indeed play in O^λ.

In general in topological models, players are not *negatively introspective*, so that there are topological models with some event E such that $u \notin \Box_i E$ and $u \notin \Box_i \neg \Box_i E$. Indeed, for many games this is the case in the model constructed for the proof of Theorem 1.5. Topological models in which negative introspection *does* hold are those in which the players' topologies are 'almost discrete'; that is, in which every open set is also a closed set; i.e. its complement is open. To get an idea of why this equivalence holds: $s \notin \Box\neg E$ iff s is in the closure of E. So if the former always implies that $s \in \Box\neg\Box\neg E$ (as negative introspection states), it means that the closure of E, $cl(E) \subseteq int(cl(E))$, which holds just if $cl(E)$ is open.

We have already mentioned that Theorem 1.5 is stated in a stronger way than usually in the literature, and we can now show that it would not be possible to prove it as stated, even in the finite case, with respect to S5 models. That is, to use the terminology of [Samet, 1990], if we were, as in the case of partitional models, to require that players do not 'ignore their ignorance', so that they could never not believe something without believing that they do not believe it, then our strong formulation of the Theorem does not in general hold. That is: it is important, for our strong formulation, that the topology not be discreet.

To see this, consider the finite game depicted in Figure 1.7. Let O_i for both players be the operator corresponding to the global version of the elimination of strategies that are strictly dominated by a pure strategy. In this game, first L is eliminated, then U, then C, then M. Now suppose towards a contradiction that there were some full partitional model (W, R_a, R_b, ξ) in which players are correct about their own strategies, and in which for all $m \in \mathbb{N}$, we have $s \in O^{1+m} \Rightarrow s \in \xi(r \cap \Box^m \mathbf{r})$. Then there would be in that model some u with $\xi(u) = (D, C) \in O^1$, in which case we would by hypothesis have $u \in \mathbf{r}$. Then we would have $D \in O_a(\xi(R_a(u)))$, which means, given the way the preferences of player a (the row player) are arranged, and that players are

	L	C	R
U	$2,0$	$0,2$	$0,1$
M	$1,0$	$1,0$	$1,1$
D	$0,0$	$0,0$	$2,1$

Figure 1.7: A game in which there is no partitional model such that for all $m \in \{0,1,2,3\}$, $O^{1+m} \subseteq \xi(\mathbf{r} \cap \Box^m \mathbf{r})$.

correct about their own strategies, that $\exists v \in R_a(u) : \xi(v) = (D, R)$. But then by 4^{r} and 5^{r}, $u \in R_a(v)$; and by hypothesis $v \in \Box^4\mathbf{r}$; in which case $u \in \Box^3\mathbf{r}$. In which case $u \in \mathbf{r} \cap \Box^3\mathbf{r}$, but that would mean (by Theorem 1.3) that $\xi(u) \in O^4$, which is clearly false. So discreet models cannot always be given with the properties of the model in Theorem 1.5.

Summary

If a strategic game is given, along with some monotonic notions of optimality, we have shown that there is then a straightforward 'epistemic foundation' for the various, possibly transfinite, rounds of iterated elimination of non-optimal strategies, in terms of some equivalent level of rationality and mutual belief of rationality.

This is a generalisation of known results in that:

- It covers arbitrary monotonic optimality operators (sometimes with a condition of globality).
- It holds for infinite games.
- Very few conditions were placed on the beliefs of the players. Notably, either
  - they need not be positively or negatively introspective, or
  - they have one introspection property mT^+, but each player might not put her various pieces of information together or draw any conclusions from them.

In getting as close as possible to a sense in which $O^{1+\alpha}$ is '*equivalent*' to $\mathbf{r} \cap B^\alpha\mathbf{r}$, we proposed a model in which the two are equivalent for all levels α. In the relational case this only works at finite levels, with all infinite levels collapsing. So we suggest that some form of neighbourhood models, specifically topological models, are appropriate for such transfinite reasoning. In topological models, a player can fail to put together only infinitarily many pieces of her information.

In the model we constructed, players were correct about their own strategies ($s_i = \Box_i s_i$ for all $s_i \in T_i$), and were positively introspective. We noted that it is *not* in

general possible to give such a model if we assume also that players are negatively introspective.

A more general point: in *all* of the models we gave, *the states of the world were just the strategy profiles*. This might seem simplistic, but also it arguably fits with the one-shot approach, in which we do not want to assume much concerning players' information. We will look in Chapter 3 at a different approach to giving an epistemic analysis of games, in which the game reduction algorithms given by the optimality operator are simulated more directly on the side of the epistemic model as 'public announcements', and the model used there will also use the strategy profiles as its state space.

Chapter 2

Syntax and Interaction

"*Le tableau, certes, est dans mon oeil.*
Mais moi, je suis dans le tableau."
– Lacan [1973]

In the previous chapter we did not explicitly introduce formal languages for reasoning about the models we discussed. That is, we did not make any distinction between syntax (languages) and semantics (models). Thus we took what has been the standard approach in the game-theoretical literature since the important early work of Aumann [1976]. As we will see, in more recent work [Aumann, 1999], the same author has argued in favour of a syntactic approach to the epistemic analysis of games. In effect this recommends the work of logicians who have studied similar models, but *started* from the syntax side, studying formal languages and axiom systems for modal logic. Aumann argues for this position, that logical syntax is important in analysing games, on the basis that the semantic approach "is conceptually not quite straightforward" [1999, p. 264], notably begging the following question: are the various parts of the model "themselves in some sense 'common knowledge'?" (p. 272). This can be seen as related to a concern raised by Brandenburger and Keisler [2006]. They see a formal language as representing in some sense the powers of representation, and formulate a condition, that they call 'completeness' of a model with respect to a language, which they interpret as meaning that the players *have access to* the language. The terms 'complete' and 'universal' are used to denote different properties in the game-theoretical literature (cf. the classification in (*ibid*, Section 11)), and have yet other connotations in the logical literature, so we will use the term 'assumption-complete' for the property described in (*ibid*). The main technical contribution of this Chapter is Theorem 2.4, that states the existence of a model that is assumption-complete for a class of modal languages. We leave open (Conjecture 2.1) the question whether the same holds for a richer language, the 'bounded fragment' of first-order logic. Assumption-completeness is, as Brandenburger and Keisler [2006] mention, related to Russell's paradox; we point out that our

Conjecture, if correct, would separate the problem of assumption-completeness from the problem of a coherent axiom of comprehension for a given language.

One advantage of the syntactic approach is that it allows us to abstract from the details of a particular model. This, crucially, allows comparisons between models of the same kind (say, two different relational models), which is a pre-requisite for doing anything interesting with logical dynamics, that we examine in Chapter 3.

Another consequence (that we do not pursue further here) is that one can make comparisons between the different kinds of model used in game theory to represent the beliefs and knowledge of players. Although we attributed Theorems from Chapter 1 to Tan and Werlang [1988], properly speaking they were working with a different kind of model. A shared syntax, with terms for belief, rationality and so on would allow for an easy precise comparison between results.

There are two important kinds of model for representing players' information: state-space models, which are the only ones that we have considered so far (relational models and neighbourhood models are both examples of state-space models), and so-called 'type-space' models, which are based on the ideas of Harsanyi [1968]. In this Chapter we will show how to translate between type-space models and a certain class of state-space models. Brandenburger and Keisler formulated the property of assumption-completeness with respect to two-player type-space models. Another contribution we make here will be to show what the notion means in our familiar state-space models.

Even disregarding the various arguments that we will look at that support separating syntax from semantics when doing formal epistemic work in games, the reader might agree that using tools from logic, with the concomitant level of abstraction that this gives us, is worthwhile in itself. Indeed, we will present (Theorem 2.1) a *syntactic* proof of Theorem 1.1, that reveals it boils down to a very simple use of the proof rules for *fixpoint operators*.

Background literature

We have mentioned as our starting point [Aumann, 1999], who presents arguments in favour of a syntactic approach. The conceptual contributions of that paper make it stand out, and technically there are a number of recent studies applying logical tools, including formal languages, to the analysis of games. We do not pretend to give an even nearly exhaustive list, but let us mention [Bonanno, 2002; Stalnaker, 1994; Benthem, 2001; Benthem *et al.*, 2006; Bruin, 2004]. The last work for example considers a *purely* syntactic approach to the epistemic analysis of games, looking at a number of solution concepts and asking what proof rules are necessary to derive them.

A number of the formal languages we consider in Section 2.2 are studied rigorously from a model-theoretic point of view in [Cate, 2005].

The question of assumption-completeness of a belief model is introduced in [Brandenburger, 2003], studied extensively in [Brandenburger and Keisler, 2006], and given further formal analysis in [Pacuit, 2007].

Organisation of the Chapter

In Section 2.1 we will define the central notions that allow us to make formal the syntax-semantics divide, and then look at the various arguments in favour of doing so. Then in Section 2.2 we catalogue a number of choices of specific language that one could make, and explore some of their properties. Notably, we look at the question of *definability* of the key notions of common belief and rationality in these languages. Section 2.3 defines type-space models and shows in what sense they correspond to (relational) state-space models. The rest of it we devote to the topic of assumption-completeness, by showing what it means on state-space models, and proving (Theorem 2.4) the existence of an assumption-complete model for some modal languages.

2.1 Features of the syntactic approach

The syntactic approach to reasoning about games involves specifying a language $\mathcal{L}$, a class of models $\mathfrak{M}$, and an interpretation (class-)function $[\![-]\!]$.

Definition 2.1. A ***language*** is a set of 'sentences' (sometimes called 'formulae') which is, in all the examples we will consider, built up recursively, according to rules of the form '$\mathbf{s}_i$ is a sentence', and 'if φ is a sentence then $\Box_i\varphi$ is a sentence'.

We have already seen some classes of models: monotonic neighbourhood models, relational models, and partitional models for games. More generally:

Definition 2.2. a ***model*** consists of a domain (in the given examples from the previous Chapter, the domain was the set of states), usually denoted W, and predicates and relations on it.

Definition 2.3. An ***interpretation function*** takes as input a sentence from the language $\varphi \in \mathcal{L}$, and a model $\mathcal{M}$, and returns an event denoted $[\![\varphi]\!]_{\mathcal{M}}$, which we will call the 'interpretation of φ in $\mathcal{M}$'.

Given a model $\mathcal{M}$, $|\mathcal{M}|$ denotes its domain. So for any formula φ, $[\![\varphi]\!]_{\mathcal{M}} \subseteq |\mathcal{M}|$.

Often when we talk of a 'language' we mean the set of sentences *plus* the interpretation function. Just as the sentences are built up recursively in the examples we consider, so will be the interpretation function.

Given some model, if the event E is the interpretation of some formula of the language $\mathcal{L}$, we say that E is ***definable*** by $\mathcal{L}$, or just '$\mathcal{L}$-definable'. Notice that this is clearly not a trivial notion, in the sense that there are models with events that for some language $\mathcal{L}$ are *not* $\mathcal{L}$-definable.

This is an essential distinction between the syntactic and the semantic approaches. The latter does not make any distinction between on the one hand 'natural' events to consider, i.e. those definable according to some language, and on the other hand arbitrary events.

We find the greatest advantage of a syntactic approach to be that it enables one to make ***inter-model*** comparisons, which is, in the spirit of [Gerbrandy and Groeneveld, 1997; Baltag *et al.*, 1999; Benthem, 1996], the best way to make sense of so-called 'dynamic' model-changing events. What we mean by inter-model comparison is as follows: Fix some formal language $\mathcal{L}$, take a formula φ from it, and then take some family of models $\{\mathcal{M}_j\}_{j\in J}$. Then *the semantics of $\mathcal{L}$ specify the meaning of φ in each model*. This might seem like a trivial point, but each formula φ has some recursively-given intended meaning, so for example in some appropriate language, $\Box_i \Box_j \mathbf{r}_i$ *means* that i believes that j believes that i is rational. Now suppose that $\{\mathcal{M}_j\}_{j\in J}$ is some tree representing changes that can be made to the epistemic situation. Then we are able to say *when* the formula φ is true.

The first example of model-changing events, that we will look at in Chapter 3, is of 'public announcement' (which we also call 'relativisation'), which is a qualitative version of Bayesian conditioning, in which the state space strictly shrinks from model to model. So for example we will look at what happens to a model when it is announced that a certain player is rational. Interesting effects are visible in the syntactic approach that would remain shrouded were we to consider a purely semantic approach. For instance, it can happen that, after a formula φ is announced, φ is no longer true, whereas this could not be captured by the semantic approach, because if the event E is true at a certain state u (i.e. $u \in E$), then as long as the event A is also true at u, then after relativising to A ('announcing' A), u will still be in (the relativised version of) E, because that will just be $E \cap A$.

Information dynamics are interesting in their own right, but we also motivate their use in game theory further in this Thesis, partly by considering (in Chapter 3) various kinds of announcements and seeing how those are related to the solution concepts explored in Chapter 1, and more substantially by giving what we argue is the correct epistemic foundation for the otherwise thorny issue of backward induction, in Chapter 4.

It is also argued in [Brandenburger and Keisler, 2006] that explicitly using a syntactic approach can offer important conceptual clarity. The authors define a property of models that they claim captures the idea of a given language $\mathcal{L}$ being "available" to player i, in the sense that if a proposition is definable in $\mathcal{L}$ then there is some state in which it defines the information that i has. This property is called 'assumption-completeness'[1], and the main Theorem of the cited work is an impossibility result, stating that for a sufficiently strong language, namely a very standard first-order language, there are *no* assumption-complete models.

One of the contributions of this Chapter is to show that the *basic modal language*, which is a standard language for reasoning about knowledge/beliefs, and is essentially that used for example in [Aumann, 1999], there *are* assumption-complete models. In fact, we show that *infinitary* modal languages have assumption-complete models. As

[1] In [Brandenburger and Keisler, 2006] it is called just 'completeness', but since we use that term in this Thesis in its more standard logical sense we prefer the less ambiguous 'assumption-completeness'.

we will separately remark in this Chapter, *infinitary* modal languages might be good candidates as languages for game theory, since they are expressive enough to be able to express the important concepts of common belief and, given some game, to define rationality.

In relatively recent work, Aumann has offered arguments in favour of taking a syntactic approach.

> "*While the semantic formalism is the more convenient and widely used of the two, it is conceptually not quite straightforward.*" – [Aumann, 1999, p. 264]

The main thrust of Aumann's argument appears to be that the syntactic approach helps to answer the question "what do the participants know about the model itself?" (op.cit., p. 272) by giving what he takes to be a more coherent account of *what a state is*. – A state u is identified with the collection of all of those formulae of the relevant language that are true at u. The idea is attributed by Aumann to Samet [1990]. This is taken to have bearing upon the troubling question of whether *the model itself* is common knowledge amongst the players.

The notion that a world can be identified with the set of sentences that it makes true is familiar from modal logic literature, on both the philosophical and technical levels. Technically, this notion is the essential ingredient to the elegant *canonical model* technique that is used to prove *completeness* of an *axiom system* with respect to some considered semantics.[2] Aumann gives a completeness result for his modal language, (cf. e.g. the textbooks [Chellas, 1980; Blackburn *et al.*, 2001]), and argues that it then becomes "clear from the construction itself that the knowledge operators are 'common knowledge' in the appropriate sense." (op.cit., p. 273; "the construction" here refers to the construction of the canonical model). The worry was that otherwise another model would have to be built on top of the given model, in order to represent uncertainty about the model itself; and of course this could continue to a vicious infinite regress. (Philosophically, the notion that a possible world *is* a set of sentences is related to a debate concerning the nature of propositions, for example between Lewis [1973] and Stalnaker [1976].)

There are connections between this idea of Aumann's that concerns whether the model can be common knowledge, and that of Brandenburger and Keisler [2006], who worry about whether the players have access to the reasoning abilities employed by the person who is building the model in order to analyse the situation they find themselves in. (And indeed, we will in effect use a canonical model construction in order to prove Theorem 2.4 giving the existence of an assumption-complete model for the modal language.)

However, Aumann does not discuss the question whether interesting relevant events are definable in the language he presents. For example, a rather natural event that one

[2]The canonical model technique was invented after Kripke's initial completeness results, that used systems of tableaux rather than a canonical model.

might want to define in the context of game theoretical analysis would be the event that a player i is rational. Another event might be common belief (or in Aumann's terms common *knowledge*) of rationality.

The language presented in [Aumann, 1999] does not include a common belief operator, nor does it include any term for rationality. Common belief is not definable in the (finitary) language in terms of belief, and indeed the modal logic completeness argument for languages that include common belief (or knowledge) is not as straightforward as for the basic modal language without such operators. It can use for example the technique of "*filtration*" [Blackburn *et al.*, 2001], and there is not one single canonical model in which sets of sentences correspond in the same meaningful way to states.[3]

Furthermore, it is not clear whether *rationality* would be definable in the modal language that Aumann considers. Of the various languages that we catalogue below, one of the aspects of each that we will consider is whether rationality is definable in it.

Of course, one could simply add to the language a symbol $\mathbf{r}_i$ that is intended to mean that player i is rational, and that is something that we will look at below in Section 2.2. However, we then lose the completeness result and the canonicity property that motivated Aumann in the first place. Perhaps we can re-establish it for the given language, but it does not come for free just from the completeness of the basic modal language.

The point of completeness is that it is about a proof system, i.e. a set of rules of syntactic manipulation of the form 'If φ is provable then ψ is provable', where in natural cases ψ is some simple syntactic manipulation of φ, that can in principle be used to determine the set of formulae that are *valid* with respect to the given semantics. That is, if a formula φ is valid, i.e. true everywhere in every model, then there is some series of legitimate syntactic manipulations, stating that φ is provable, i.e. a *proof* of φ. We will call the set of sentences that are valid with respect to a given class of models the 'logic' of that class of models (with respect to some given language), and we call the elements of a logic its 'theorems'. If the language is rich enough, then the syntactic proof system itself could be used for proving theorems. (It can also facilitate checking which assumptions are actually needed. Indeed, that is how we arrived at the minimal requirements for Theorem 1.1, and we will provide a simple syntactic proof of that theorem below in Section 2.2.)

[3] This is related to the fact that the resulting logic is not *compact*, meaning that there can be an infinite set of sentences Γ that is not satisfiable (cannot be true) but such that each finite subset of it *is* satisfiable (*can* be true). In this case the standard argument would for example set:

$$\Gamma = \{\Diamond^* \neg p\} \cup \{\underbrace{\Box \ldots \Box}_{m \text{ times}} p \mid m \in \mathbb{N}\}.$$

2.2 Languages

In this Section we will catalogue a number of different formal languages with expressions that will stand for aspects of the various items in the state-space models that we have considered. This means specifying the set of sentences and the interpretation function. All of the languages we consider will be recursively defined. We use standard compact representations for these recursively-defined objects. To give a set of sentences, we write expressions of the following BackusNaur form:

$$\varphi ::= a_1 \mid a_2 \ldots \mid f_1(\varphi, \ldots, \varphi) \mid \ldots \mid f_m(\varphi, \ldots, \varphi)$$

This should be read as saying that $\mathcal{L}$, the language in question, is defined as the smallest set that:

1. contains each of $a_1, a_2, \ldots$; and
2. if it contains $\varphi_1, \ldots, \varphi_k$ then it contains $f(\varphi_1, \ldots, \varphi_k)$.

For specifying the semantics (i.e. the interpretation), we will use both the $[\![-]\!]$ notation, and also sometimes write $u \vDash_{\mathcal{M}} \varphi$ to mean $u \in [\![\varphi]\!]_{\mathcal{M}}$. As with the $[\![-]\!]$ notation, for notational elegance we sometimes drop the symbol $\mathcal{M}$ for the model when the model is clear from the context.

For reference purposes we give a summary table of the various language elements at the end of this Chapter.

In the modal logic literature, one often considers a slightly more general class of models than those given in the previous Chapter. There due to our specific aim we only considered models for specific games G, where the only non-epistemic fact associated with each state was a strategy profile from the game G. (And indeed, we discussed the possibility of only considering models in which the state space *is* the set of strategy profiles.) In this Chapter we still do not allow that other 'atomic' (non-epistemic) information than just the strategy profiles can be specified for a particular state. However, this particular decision does not result in the loss of any generality: what we say will for the most part also hold for the more general modal models, in which a model would be parametrised by an alphabet of 'atomic propositions' Ψ, so that a Ψ-model would be a structure $(W, \mathcal{N}, V)_{i \in N}$, with $V : \Psi \to 2^W$ a ***valuation function*** assigning an event[4] to each of the atomic propositions. These general models are the ones that are considered in [Aumann, 1976] and in the modal logic literature (see e.g. [Chellas, 1980]).

The models we consider are instances of these more general models, in which for example we set $\Psi = \bigcup_{i \in N} S_i$, so one option for the language, is to have an atomic proposition for each one of each player's strategies.

Notice though that for *general* modal models to be faithful to the idea of what a strategy is supposed to represent, we also would need to impose the restriction that

[4] So implicitly we're in the context of some model.

each player chooses precisely *one* strategy at any given state. That is, we would need to say that for every player i and every state u, there is an $s_i \in S_i$ such that $u \in V(s_i)$; and that if $u \in V(s_i)$ with $s_i \in S_i$, then for all $s_i' \in S_i - \{s_i\}$, $u \notin V(s_i')$. So the $V(s_i)$ would have to form a partition of the state space W.

Another way to represent the players' strategy choices in the language is by using a modality, writing for example $[i_s]\varphi$ with the intended meaning, 'in all states where i plays the same strategy, φ holds'. Since, as we just noted, the strategy choices form a partition of the space, this would be an $S5$ modality. This is in effect the choice taken in [Benthem, 2007b], where the author specifies that the epistemic relation of each player i in the model is determined precisely by saying that at state u she considers state v plausible just if she chooses the same strategy at both states. (We discuss that work further in Section 3.2.)

Since we are interested in more general epistemic relations, it makes sense for us to consider introducing this separate modality $[i_s]$ in addition to $\Box_i$, and we will consider such languages briefly in what follows.

Another option is to have proposition letters for *outcomes* rather than individual strategies. This will be more natural in Chapter 4, when we look at *extensive games*, and want to talk about outcomes rather than strategies We will not take strategies as primitive there, since in extensive games strategies are more complex and conceptually loaded with counterfactuals. In the case of strategic games it might seem less natural, though it will be easy enough, at least in the case of finite strategy sets and players, to define strategies in terms of strategy profiles, and in the case of finite players to define strategy profiles in terms of strategies.

Once we have specified language, we will be interested in the question whether it can define some subset or operator on the model. For a language to be able to define an event, for example the event of rationality, means that there must be some formula φ in the language such that in every model $\mathcal{M}$, $[\![\varphi]\!]_{\mathcal{M}}$ is the event in question. Similarly, take some unary operator $F : 2^W \to 2^W$ on the model. For example, common belief would be the operator that takes an event $E \subseteq W$ and returns the event that E is common belief. Define a *unary formula-scheme* to be a formula except that it has a place-holder for another formula. For example, $\varphi(\psi) := \psi \wedge s_i$ is a definition of a formula-scheme. Then for a language to be able to define the unary operator F means that there is some unary formula-scheme $\varphi(\psi)$ in the language such that in every model $\mathcal{M}$, $F([\![\psi]\!]_{\mathcal{M}}) = [\![\varphi(\psi)]\!]_{\mathcal{M}}$.

There is another sort of definability, which we will not be able to cover in any detail in this Chapter, that we call 'axiom definability'. For a language to be able to axiomatically define an event (operator) means that if we were to add some proposition letter (modality) to the language, but not to specify its interpretation, then just by stating certain validities ('axioms') in this new language, and thereby restricting the class of models that are allowed, one can force the desired interpretation, so that the proposition (modality) is always interpreted by the event (operator). We do not devote much time to questions of axiomatics in this Chapter. Axiom definability is not only studied in the logic literature, but is also considered by formal epistemologists [Halpern *et al.*,

2007].

In general when we talk about 'definability' we will mean the first kind of definability, and sometimes we say that a language can 'express' something to mean that it can define the relevant event/operator.

The 'basic modal language' that we first present is standard from the modal logic literature [Blackburn *et al.*, 2001]. It is also essentially the modal language considered in [Aumann, 1999]. The basic modal language is parametrised by a set of 'atomic propositions' Ψ and a set of players N. This could in principle be anything, but in the examples we are considering, as discussed above, it will be some subset of $T \cup \bigcup_{i \in N} T_i \cup \{\top\}$, i.e. the set of strategy profiles, the set of strategies and some constant that will stand for 'truth'.

Then the sentences of the ***basic (finitary) modal language*** $\mathcal{L}_N^\Psi$ are given as follows, where $p \in \Psi$ and $i \in N$:

$$\varphi ::= p \mid \neg\varphi \mid (\varphi \wedge \varphi) \mid \Box_i \varphi$$

The interpretation of $\mathcal{L}_N^\Psi$ is specified with respect to a model $\mathcal{M}_G = (W, \mathcal{N}_i, \xi)_{i \in N}$ of the game G as follows:

$$\begin{array}{lll}
\mathcal{M}_G, u \vDash \top & & \\
\mathcal{M}_G, u \vDash s_i & \text{iff} & \xi_i(u) = s_i \\
\mathcal{M}_G, u \vDash s & \text{iff} & \xi(u) = s \\
\mathcal{M}_G, u \vDash (\phi \wedge \psi) & \text{iff} & u \in [\![\varphi]\!] \cap [\![\psi]\!] \\
\mathcal{M}_G, u \vDash \neg\varphi & \text{iff} & u \notin [\![\varphi]\!] \\
\mathcal{M}_G, u \vDash \Box_i \varphi & \text{iff} & R_i(u) \subseteq [\![\varphi]\!]
\end{array}$$

We sometimes don't write brackets when they're not necessary for disambiguation, and we use many standard abbreviations from propositional logic, writing $\varphi \vee \psi$ for $\neg(\neg\varphi \wedge \neg\psi)$, $\varphi \to \psi$ for $\neg(\varphi \wedge \neg\psi)$ and $\varphi \equiv \psi$ for $(\varphi \to \psi) \wedge (\psi \to \varphi)$. We also use a standard abbreviation in modal logic, writing $\Diamond_i$ for $\neg \Box_i \neg$. Where $\Box_i \varphi$ means that i believes that φ, $\Diamond_i \varphi$ means that player i does not believe that $\neg\varphi$, i.e. has not ruled out the possibility that φ, or to put it otherwise considers it possible (or 'plausible') that φ. The semantic clause for $\Box_i$ yields the following version for $\Diamond_i$:

$$\mathcal{M}_G, u \vDash \Diamond_i \varphi \quad \text{iff} \quad R_i(u) \cap [\![\varphi]\!] \neq \emptyset$$

A number of notions that we consider involve non-finite things. One example of this is the notion of *common knowledge*. However, we will also see that in the case of arbitrary games (so with possibly infinite strategy sets), it will be convenient, in order to define other events, for example the important event of a player's being *rational*, to consider languages with infinitely long expressions.

The kind of infinite languages we consider are those with conjunctions or disjunctions taken over infinitely many sentences. Notice that the basic finitary modal language allows for conjunctions of arbitrary finite length, meaning that for any *finite* set

$\Phi = \{\varphi_1, \ldots, \varphi_k\}$ of sentences, $\bigwedge \Phi$ is a sentence, where $\bigwedge \Phi = \varphi_1 \wedge \varphi_2 \wedge \ldots \wedge \varphi_k$. However, there are instances of the finitary modal language that contain (infinite) sets of sentences Φ, and there is no single sentence that is equivalent to the conjunction $\bigwedge \Phi$ of all those sentences.

The size of a language is specified in terms of its *cardinality*. For any set X, we write $\#(X)$ to denote the ***cardinality*** of X. A cardinal κ is called ***inaccessible*** just if $\alpha < \kappa \Rightarrow 2^\alpha < \kappa$. Note that the first infinite cardinal, denoted $\aleph_0$, is inaccessible, since for any finite n, 2^n is also finite.

Given some infinite cardinal κ, the ***basic infinitary modal language*** of cardinality κ, $\mathcal{L}^{\Psi}_{N,\kappa}$, is defined as follows:

$$\varphi ::= p \mid \neg\varphi \mid \bigwedge \Phi \mid \Box_i \varphi,$$

where $p \in \Psi$ and Φ is a subset of sentences of $\mathcal{L}^{\Psi}_{N,\kappa}$, of cardinality strictly less than κ. (Remember that this notation is shorthand for a recursive definition, so this is not circular.) Semantically we give the following natural interpretation to infinite conjunctions:

$$\mathcal{M}_G, u \vDash \bigwedge \Phi \quad \text{iff} \quad \text{for all } \varphi \in \Phi,\ \mathcal{M}_G, u \vDash \varphi.$$

Thus notice that the basic finitary modal language $\mathcal{L}^{\Psi}_{N}$ is just (a notational variant of) $\mathcal{L}^{\Psi}_{N,\aleph_0}$. We often write $\phi \wedge \psi$ for $\bigwedge\{\phi, \psi\}$, and $\bigvee \Phi$ is an abbreviation of $\neg \bigwedge\{\neg\varphi \mid \varphi \in \Phi\}$.

Another modality that we might consider adding is a so-called 'global' or 'universal' modality. (It is important to notice that this sense of 'global' is not related to the term we use to distinguish between different versions of optimality operators.) The global modality does not look at the current state, but looks rather at the whole model to decide whether the statement it expresses is true. We write it $A\varphi$, so the language $\mathcal{L}_A$ with a universal modality is formed from the language $\mathcal{L}$ by adding the following clause:

$$\varphi ::= \ldots \mid A\varphi,$$

and it has the following semantics:

$$\mathcal{M}_G, u \vDash A\varphi \quad \text{iff} \quad [\![\varphi]\!]_{\mathcal{M}} = W.$$

This 'modality' does not respect what we suggest is the key property of modal languages: their *locality*. In order to evaluate whether a given formula of a modal language is true at some state u, one need not look at states that are not 'related' to u,

As we said, we could also introduce for each player i modality $[i_s]\varphi$, that will look, in order to determine whether it is true at some state u, only at states where the player plays the same strategy as at u, and check whether φ holds there. This would mean the language $\mathcal{L}$ could be extended to $\mathcal{L}_{i_s}$ as follows:

$$\varphi ::= \ldots \mid [i_s]\varphi,$$

and the following semantic clause would give the meaning of $[i_s]$:

$$\mathcal{M}, u \vDash [i_s]\varphi \quad \text{iff} \quad \text{for all } v\, \xi_i(u) = \xi_i(v) \Rightarrow \mathcal{M}, u \vDash \varphi$$

So this modality in effect introduces a new relation into the model: two states become related to each other if the player plays the same strategy.

Notice then that the global modality A is definable, at least in case there are a finite number of players. Let $[A_s]\varphi$ be defined as follows, where w is a list of the players N:

$$[A_s]\varphi := [w_{0s}][w_{1s}] \dots [w_{\#(N)-1_s}]\varphi.$$

Then Fact 2.1 states that $[A_s]$ defines the global modality.

Fact 2.1. *On any model $\mathcal{M}$,*

$$[\![[A_s]\varphi]\!]_{\mathcal{M}} = \begin{cases} W & \text{if } [\![\varphi]\!]_{\mathcal{M}} = W \\ \emptyset & \text{otherwise.} \end{cases}$$

Let us look at the notion of 'locality' that we have mentioned. Now, historically speaking locality is *not* what modality was about in the case of relational semantics: early studies concerned $S5$ in the case of a single modality, in which case the modality is thought of as a global modality. However, the following quotation shows that contemporarily the same is not true:

> "*Modal languages provide an internal, local perspective on relational structures.*" – [Blackburn *et al.*, 2001]

But more importantly than whether a given language *is*, according to some mysterious essentialist classification, modal, are two more serious points, both concerning the notion of whether the language is 'available' to the players, for them to use to think about the model they are in.

The first we can only express informally, by saying that we find that a global modality transcends the appealing notion that what is considered possible by the players is really given by the relations in the case of a relational model.

The second point is a little more formal: we will see later that adding a global modality to an otherwise 'local' language breaks the property of assumption-completeness. However, this point is also made more flimsy by the fact that the particular local language considered might itself not have the property of assumption-completeness.

Clearly neither of these points speak conclusively against including a global modality in a language! Indeed, both of these arguments are tenuous, and furthermore both are predicated on the notion that the language is meant to capture some sense of what the players can represent to themselves. Yet as we have seen at the beginning of this Chapter, there are many other reasons to introduce a formal syntax. Nonetheless, we do find the locality idea appealing, and take even these two partially-formed and inconclusive points to motivate focusing on more local languages.

As a final point for consideration, note that if one accepts Aumann's arguments about the importance of the existence of a single canonical structure for a given language, then non-local languages, at least any language in which one can define a global modality, would also not be acceptable. This is because there is no canonical model for a language with a global modality.[5]

The next addition to this basic modal language that we consider is adding optimality operators. The language $\mathcal{L}_O$ is the language $\mathcal{L}$ along with the following clause, where $i \in N$:

$$\varphi ::= \ldots \mid \bigcirc_i \varphi.$$

The most general interpretation we could give to this operator would be that it is an arbitrary *monotonic* operator. This could be formulated in a number of different ways, but essentially means that the optimality operator has the same *monotonic neighbourhood* semantics as we saw before. That is, we would interpret it on a structure $(W, \mathcal{N}_i, O_i, \ldots)_{i \in N}$ where each O_i, like each $\mathcal{N}_i$, is a monotonic neighbourhood function. (The definition of monotonic neighbourhood models was given above in Section 1.4.)

Since rationality was defined purely in terms of information and optimality, and we can express both of these things in this language, the question of whether rationality is definable in the language is now not entirely trivial. If we were to follow the suggestion we have just seen, and have a semantics in which there are no constraints placed on the optimality operator, then the basic modal language would not be able to define rationality.

Proposition 2.1. *Suppose that the semantics were to interpret the O_i as arbitrary monotonic operators on the state space. Then there would be no $\mathcal{L}^{\{\top\}}_{N,O}$-formula φ such that $\omega \vDash \varphi \Leftrightarrow \omega \in \mathbf{r}$.*

Proof. To prove this proposition, we use the notion of $\mathcal{L}^{\Psi}_{N,O}$-***bisimulation***.

Definition 2.4. An $\mathcal{L}^{\Psi}_{N,O}$-bisimulation between two relational models with monotonic optimality operators $\mathcal{M} = (W, R_i, O_i, \xi)_{i \in N}$ and $\mathcal{M}' = (W', R'_i, O', \xi')_{i \in N}$ is a relation $Z \subseteq W \times W'$ satisfying the following conditions:

1. $uZu' \Rightarrow (u \vDash p \Leftrightarrow u' \vDash p)$, for $p \in \Psi$;
2. $uZu' \Rightarrow (w \in R_i(u) \Rightarrow \exists w' \in R'_i(u') : wZw')$;
3. $uZu' \Rightarrow (u \in O_i(X) \Rightarrow \exists X' \subseteq W' : u' \in O'_i(X') \;\&\; \forall x' \in X' \exists x \in X : xZx')$;
4. $uZu' \Rightarrow (w' \in R'_i(u') \Rightarrow \exists w \in R_i(u) : wZw')$;

[5] That is, the 'truth lemma' $\Gamma \vDash \varphi \Leftrightarrow \varphi \in \Gamma$ cannot hold in any model containing as states the maximally consistent (or satisfiable) sets of formulae, since both Ap and $A\neg p$ are consistent (satisfiable), so there would be Γ, Γ' in the model with $Ap \in \Gamma$ and $A\neg p$ in the model, but it conflicts with the semantics of A to have in the same model $\Gamma \vDash Ap$ and $\Gamma' \vDash A\neg p$.

5. $uZu' \Rightarrow (u' \in O'_i(X') \Rightarrow \exists X \subseteq W : u \in O_i(X)\ \&\ \forall x \in X \exists x' \in X' : xZx')$.

This kind of bisimulation has the following important property:

Proposition 2.2. *For any $\mathcal{L}^{\Psi}_{N,O}$-bisimulation Z between $\mathcal{M}$ and $\mathcal{M}'$, any ω, ω' such that $\omega Z \omega'$, and any formula $\varphi \in \mathcal{L}^{\Psi}_{N,O}$, the following equivalence holds:*

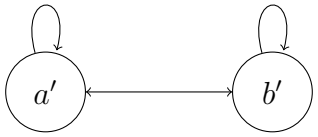
$$\mathcal{M}, \omega \vDash \varphi \Leftrightarrow \mathcal{M}', \omega' \vDash \varphi$$

Proof. By induction on the formula φ; cf. [Blackburn *et al.*, 2001] and [Hansen, 2003, Proposition 4.10]. ■

Then to prove Proposition 2.1, consider the two models $\mathcal{M}$ and $\mathcal{M}'$, given in Figures 2.1 and 2.2. (In these models, we draw the relation for just one player, and list the values for the monotonic operator that would be interpreted as her optimality operator.)

$$O(\{a\}) = \{b\}; O(\{b\}) = O(\{a, b\}) = \{a, b\}$$

Figure 2.1: The model $\mathcal{M}$ used to prove Proposition 2.1.



$$O(\{a'\}) = \{b'\}; O(\{b'\}) = O(\{a', b'\}) = \{a', b'\}$$

Figure 2.2: The model $\mathcal{M}'$ used in proving Proposition 2.1.

$Z = \{a, b\} \times \{a', b'\}$ is an $\mathcal{L}^{\top}_{N,O}$-bisimulation between $\mathcal{M}$ and $\mathcal{M}$'. Then suppose towards a contradiction that there *is* a formula $\rho_i \in \mathcal{L}^{\top}_{N,O}$ that defines rationality of i, i.e. such that $\omega \vDash \rho_i \Leftrightarrow \omega \in O_i(R_i(\omega))$. Then we would have $a \nvDash \rho_i$, since $a \notin O_i(R_i(a)) = O_i(\{a\}) = \{b\})$, and $a' \vDash \rho_i$, as $a' \in_I (R_i(a')) = O_i(\{a, b\}) = \{a, b\}$. But this would contradict Proposition 2.2, because aZa'. ■

Notice that the language we just considered did *not* have terms for the strategies or strategy profiles of the players. Still, we could easily ensure that the same result *does*

hold for a language with such terms s_i or s, simply by saying that at a and b in the model, the players play the same strategies. However, that would go against the sense of what an optimality operator should be.

Indeed, rather than interpreting $\bigcirc_i$ as some *arbitrary monotonic operator*, it would be more accurate to interpret it really as an *optimality operator*. So in particular, two conditions we will want to impose are the following:

1. If $\xi_i(u) = \xi_i(v)$ then $u \vDash \bigcirc_i\varphi \Leftrightarrow v \vDash \bigcirc_i\varphi$;
2. If $\xi([\![\varphi]\!]) = \xi([\![\psi]\!])$ then $u \vDash \bigcirc\varphi \equiv \bigcirc\psi$.

That is: (1) if player i plays *the same thing* as u and v then she plays optimally with respect to the restriction defined by φ at u iff she plays optimality with respect to the restriction defined by φ at v. And 2 is just the truism that if the restriction defined by φ is the same as the restriction defined by ψ then the players play optimally with respect to the restriction defined by φ just if they play optimally with respect to that defined by ψ. Note that if we are considering *global* operators then we could refine 2 further as follows:

2'. If $\xi_{-i}([\![\varphi]\!]) = \xi_{-i}([\![\psi]\!])$ then $u \vDash \bigcirc_i\varphi \equiv \bigcirc_i\psi$.

Clearly 2' entails 2, but is strictly stronger than it. It is unclear whether there are further restrictions that must be placed on an abstract monotonic operator in order for it to count as an *optimality* operator, but the conditions presented above are certainly necessary, if for some particular game and optimality operator the following semantic clause interprets the syntactic optimality operator:

$$\mathcal{M}_G, u \vDash \bigcirc_i\varphi \quad \text{iff} \quad \xi_i(u) \in O_i(\xi[[\varphi]]).$$

We therefore restrict our attention in what follows, when looking at questions of definability, to the class of models in which the two conditions above hold.

One way to reason about optimality operators, or rationality, is to define them explicitly in the language. So rather than introduce an operator $\bigcirc_i\varphi$ meaning that i plays optimally with respect to the restriction defined by φ, we might want a language which for any formula φ has for example some formula $\text{NSD}_i(\varphi)$ that expresses that player i plays a strategy that is not strictly dominated amongst the restriction defined by φ.

In order to do that the language would need some way to talk about the *preferences* of the players. We will consider here ordinal preferences (inspired by a language with terms for cardinal preferences proposed in [Bruin, 2004]), and as with strategies, we will look at two ways of reasoning about them in the language: using proposition letters or modalities.

One way to express preferences over outcomes is to use expressions of the form $\mathbf{s} <_i \mathbf{s}'$, which we interpret as meaning that i prefers (strictly) the outcome s' to the outcome s. So if $\mathcal{L}^\Psi$ has a proposition $\mathbf{s} <_i \mathbf{s}' \in \Psi$ for every $s, s' \in T$, that are interpreted in the appropriate way, we say that it has '*propositions for preferences*'.

The 'appropriate way' is just that we want $\mathbf{s} <_i \mathbf{s}'$ to be true just if $s <_i s'$. Notice that this means the valuation of each $\mathbf{s} <_i \mathbf{s}'$ is going to be everywhere true or everywhere false.

The other way to express preferences is in terms of preference modalities. [Benthem *et al.*, 2006] show how, in a suitably rich 'hybrid' logical language – effectively the language with $\downarrow$ that we define below – it is possible to define the important game-theoretical notion of Nash equilibrium. In this Chapter we do not consider solution concepts that require anything other than a deductive interpretation[6], but as we shall see, preference modalities can be used, as can preference propositions, to express some optimality and rationality notions.

Several different preference modalities are possible, and each of the following four clauses of a language $\mathcal{L}$ could be added to form, respectively, the languages $\mathcal{L}_{\geq}$, $\mathcal{L}_{>}$, $\mathcal{L}_{\leq}$, and $\mathcal{L}_{<}$:

$$\varphi ::= \ldots \mid \langle \geq_i \rangle \varphi \mid \langle >_i \rangle \varphi \mid \langle \leq_i \rangle \varphi \mid \langle <_i \rangle \varphi$$

These modalities have a natural interpretation; we give that for $\langle \geq \rangle$, the rest being analogous:

$$\mathcal{M}, u \vDash \langle \geq_i \rangle \varphi \quad \text{iff} \quad \exists s_i \in \xi_i([\![\varphi]\!]_{\mathcal{M}}) : \xi_i(u) \geq_i s_i$$

Notice then that in the language $\mathcal{L}_{\geq,\leq}$, the universal modality becomes definable: since the preference order is assumed to be total, $\langle \geq \rangle \varphi \vee \langle \leq \rangle \varphi$ is true just if $[\![\varphi]\!] \neq \emptyset$.

In any case, we will see that we need a global modality in order to define the optimality operator corresponding to non-strict dominance. The same will not, perhaps surprisingly, be true for defining rationality, which therefore remains in some sense a 'local' property.

Either of these approaches can be used, in languages with propositions for strategies, and sufficient cardinality, to express both non-strict dominance, and the corresponding form of rationality.

In order to express non-strict dominance we need the global modality (or, as we have seen, enough strategy or preference modalities). Then if we have preference propositions in the language, we define the formula scheme $\text{NSD}(\varphi)$ as follows:

$$\text{NSD}_i(\varphi) := \bigwedge_{u_i \in T_i} \Big(u_i \rightarrow \bigwedge_{t_i \in T_i} \bigvee_{s_{-i} \in T_i} \Big(E(s_i \wedge \varphi) \wedge \neg(\mathbf{u_i}, \mathbf{s_{-i}}) <_i (\mathbf{t_i}, \mathbf{s_{-i}}) \Big) \Big).$$

A very similar formula in $\mathcal{L}_{N,A,\geq}$ can also be used:

$$\text{NSD}'_i(\varphi) := \bigwedge_{u_i \in T_i} \Big(u_i \rightarrow \bigwedge_{t_i \in T_i} \bigvee_{s_{-i} \in T_i} \Big(E((s_i \wedge \varphi) \wedge \langle \geq_i \rangle (\mathbf{t_i} \wedge \mathbf{s_{-i}})) \Big) \Big).$$

Both of these formulae work to define the relevant optimality operator: Fact 2.2 states that both $\text{NSD}'_i(\varphi)$ and $\text{NSD}_i(\varphi)$ are true just when player i chooses a strategy that is not strictly dominated with respect to the restriction defined by φ.

[6] We do mention Nash equilibrium in Chapter 4; see Definition 4.7

Fact 2.2. $\xi_i([\![\text{NSD}'_i(\varphi)]\!]) = \xi_i([\![\text{NSD}_i(\varphi)]\!]) = \{s_i \in T_i \mid nsd(s_i, T_i, \xi_{-i}([\![\varphi]\!]))\}$

Being a potential best response which, recall from Chapter 1, is a defined by switching two quantifiers in the definition of not being strictly dominated, is definable by switching the second conjunction and the following disjunction. And with more preference modalities, or just using the preference propositions, note that it is also possible using this approach to define weak dominance ('admissibility'), for example as:

$$\text{NWD}_i(\varphi) := \bigwedge_{u_i \in T_i} \Big(u_i \to \bigwedge_{t_i \in T_i} \Big(\bigvee_{s_{-i} \in T_i} (E(s_{-i} \wedge \varphi) \wedge (\mathbf{t_i}, \mathbf{s_{-i}}) <_i (\mathbf{u_i}, \mathbf{s_{-i}})) \\ \vee \bigwedge_{s_{-i} \in T_i} (E(s_{-i} \wedge \varphi) \to \neg(\mathbf{u_i}, \mathbf{s_{-i}}) <_i (\mathbf{t_i}, \mathbf{s_{-i}})) \Big) \Big)$$

The syntactic form of $\text{NSD}(\varphi)$ reveals that, as we mentioned in Chapter 1, since it is indeed positive in φ, it corresponds to a monotonic operator. What we mean by that is that if $[\![\varphi]\!] \subseteq [\![\psi]\!]$ then $[\![\text{NSD}(\varphi)]\!] \subseteq [\![\text{NSD}(\psi)]\!]$. The same is not true of $\text{NWD}(\varphi)$: the second occurrence of φ occurs in a 'negative' position, in fact under the scope of a single negation symbol (recall that $\varphi \to \psi$ is an abbreviation of $\neg(\varphi \wedge \neg\psi)$).

Interestingly, in the relational case, to define *rationality* in this way, we do not need a global modality. Let us again focus on the case of non-strict dominance, so that being rational means playing a strategy that is not strictly dominated with respect to the restriction defined by your information relation. Then the following sentence expresses that player i is rational:

$$\bigwedge_{u_i \in T_i} \Big(u_i \to \bigwedge_{t_i \in T_i} \bigvee_{s_{-i} \in T_i} \Big(\Diamond_i (s_{-i} \wedge \varphi) \wedge \neg(\mathbf{u_i}, \mathbf{s_{-i}}) <_i (\mathbf{t_i}, \mathbf{s_{-i}}) \Big) \Big).$$

Similar versions can also be given for other optimality notions.

So it can be possible to define rationality if the language is able to reason fully about the strategies. What if the language is unable to express the strategies of the players? It is still sensible to ask for languages that *can* define rationality even when they lack proposition letters for strategies, or if they do not express preferences directly, and the semantic restrictions (1) or (2) were not placed on the optimality operator.

The first language we will look at in which rationality is definable is a very expressive extension of the basic modal language, in which we allow for 'second-order' quantifiers over events. This means introducing a set of proposition letters that will be used as variables VAR, written as $X, Y, \ldots$, and if $\mathcal{L}$ is a modal language then the ***propositionally quantified language*** $\mathcal{L}_{\overline{\forall}}$ based in $\mathcal{L}$ is given by taking the recursive definition of $\mathcal{L}$ and adding the following two clauses:

$$\varphi ::= \ldots \mid X \mid \overline{\forall} X.\varphi,$$

where X is one of the variables. We also use the standard abbreviation $\overline{\exists} X.\varphi$ for $\neg\overline{\forall} X.\neg\varphi$. In order to give the recursive definition of the semantics of this language,

we now have to keep track of the meaning of the variables VAR. This is done using second-order 'assignments' δ, which are functions assigning an event to each variable.

The semantic clauses are then given as a relation between a model a state *an assignment* and a formula, and are written as follows:

$$\mathcal{M}, u \vDash_\delta \varphi.$$

Similarly, we write $[\![\varphi]\!]^\delta_{\mathcal{M}}$ for

$$\{u \in W \mid \mathcal{M}, u \vDash_\delta \varphi\}.$$

The assignment is effectively 'ignored' by the existing clauses, so we would simply re-write them as they are with the subscript, for example:

$$\mathcal{M}, u \vDash_\delta \neg\varphi \quad \text{iff} \quad u \notin [\![\varphi]\!]^\delta_{\mathcal{M}}.$$

Furthermore, $\mathcal{M}, u \vDash \varphi$, where no assignment is specified, is shorthand for saying that $\mathcal{M}, u \vDash_\delta \varphi$ for *all* assignments δ. ***Closed*** formulae are those in which no variable X occurs that is not within the scope of a quantifier $\overline{\forall}X$; clearly for any closed formula φ, it is equivalent to write $\mathcal{M}, u \vDash \varphi$ or $\mathcal{M}, u \vDash_\delta \varphi$ for some (arbitrary) assignment δ.

Given a variable X and two assignments δ and δ', we write $\delta \sim_{-X} \delta'$ to mean that δ and δ' agree on all variables except (possibly) for X. I.e. if δ and δ' are assignments, then

$$\delta \sim_{-X} \delta' \text{ iff } \forall Y \in \text{VAR}(X \neq Y \Rightarrow \delta(Y) = \delta'(Y)).$$

The new semantic clauses can now be given as follows:

$$\begin{array}{lll} \mathcal{M}, u \vDash_\delta X & \text{iff} & u \in \delta(X) \\ \mathcal{M}, u \vDash_\delta \overline{\forall}X\varphi & \text{iff} & \text{for all } \delta' \sim_{-X} \delta,\ \mathcal{M}, u \vDash_{\delta'} \varphi. \end{array}$$

It is then straightforward to see that rationality is definable in any modal language with monotonic operators for optimality that is supplemented with second-order quantifiers. That is what is stated in Fact 2.3

Fact 2.3. *In any language that extends $\mathcal{L}^{\emptyset}_{N,O,\overline{\forall}}$, the event $\mathbf{r}_i$ that player i is rational is definable.*

Proof. The sentence $\overline{\forall}X(\Box_i X \to \bigcirc_i X)$ defines the event that i is rational: it says that for any event X that i believes to be true, i plays optimally with respect to X. ■

Second-order modal languages are very expressive. In fact, all of the possible quantifiers and operators that we will discuss below can be expressed in terms of second-order quantification. Second-order quantification was studied in modal languages by Fine [1970], who shows undecidability with respect to the general semantics given here (he also shows that under certain conditions on what sets can be quantified over the logic can be better behaved).

Second-order quantification is certainly not 'local' in our sense: on the contrary it 'looks at' *every* subset of the model.

In the context of relational models, there is another natural addition to the language that one can make, which is to introduce a modality corresponding to the *inverse relation* for each player. Given a relation $R \subseteq W \times W$, its ***inverse***, written R^{-1} is defined as:

$$xR^{-1}y \text{ iff } yRx.$$

Then the modality corresponding to the inverse of i's relation, which can write $\Box_i^{-1}$, has the following semantics:

$$\mathcal{M}, u \vDash \Box_i^{-1}\varphi \quad \text{iff} \quad R_i^{-1}(u) \subseteq [\![\varphi]\!]_{\mathcal{M}}.$$

Intuitively speaking it is not quite clear how to read $\Box_i^{-1}$, but perhaps the dual $\Diamond_i^{-1}$ is easier: that is true just when there is a state where φ holds, and at which i considers the current state to be possible.

Mathematically speaking, we can also think of the inverse modality as internalising the relation R_i into the language, in the sense made clear by Fact 2.4.

Fact 2.4. $[\![\Diamond_i^{-1}\varphi]\!] = R_i([\![\varphi]\!])$.

Of course, in models where the players' relations R_i are *symmetric*, for example in S5 (partitional) models, $\Box_i^{-1}$ is equivalent to $\Box_i$, and so the expressivity of the language would remain the same if we were to add these 'inverse' operators. In general though, that is not true. In $S4$ models, adding the inverse modality will add expressivity, and indeed an $S5$ operator can be defined there using the inverse modality: if we think of $\Box\varphi \wedge \Box^{-1}\varphi$ as a single modal operator with the argument φ, then of course if $\Box$ is $S4$ then the conjunction will be $S5$.

For any language $\mathcal{L}$, we write the language with inverse belief modalities as $\mathcal{L}_{-1}$.

We can use Fact 2.4 in order to define rationality in the context of the next kind of language we consider. Hybrid languages are modal but can also include constants variables and various forms of quantification over points. Hybrid constants are like standard propositional variables, except that they are true at precisely *one* state, so they are names for what we might call 'point events'. These hybrid constants are called 'nominals', and can be used to increase the axiom definability of a language. Hybrid logics of all the kinds that we consider below have been studied model-theoretically and axiomatically in [Cate, 2005].

We will not consider languages with non-variable nominals here, but will consider only languages with quantifiers over states. The first kind of quantifier that we consider is the standard first-order quantifier $\forall$. To add it to a language involves adding two clauses reminiscent of those added to form a propositionally quantified language: we again use a set of variables VAR, and a quantifier $\forall$, so we add the following two clauses, to form $\mathcal{L}_\forall$ from $\mathcal{L}$:

$$\varphi ::= \ldots \mid x \mid \forall x.\varphi,$$

where $x \in \text{VAR}$. The semantic clauses also need to keep track of an assignment:

$$\mathcal{M}, u \vDash_\delta \forall x.\varphi \quad \text{iff} \quad \text{for all } \delta' \sim_{-x} \delta,\ \mathcal{M}, u \vDash_{\delta'} \varphi.$$

Note that while $\mathcal{L}_{\overline{\forall}}$ cannot express the fact that a given event is a singleton event, and so is not as expressive as $\mathcal{L}_\forall$, $\mathcal{L}_{\overline{\forall},A}$ *can* express this fact, and so is strictly more expressive than $\mathcal{L}_\forall$.

Another operator that is often considered in hybrid languages is the $@_x$ modality, where $@_x\varphi$ means that 'at state x, φ holds.' The language clause, to form $\mathcal{L}_@$ from the hybrid language $\mathcal{L}$, given some set VAR of variables, is the following:

$$\varphi ::= \ldots \mid @_x\varphi,$$

where x is a variable from VAR. The semantic clause is then precisely as we said:

$$\mathcal{M}, u \vDash_\delta @_x\varphi \quad \text{iff} \quad \delta(x) = \{v\}\ \&\ \mathcal{M}, v \vDash_\delta \varphi.$$

Notice that if the universal modality were added to a hybrid language, then that would already be able to express the $@_x$ modality, since in the case where x can only be interpreted as singleton or empty events, the sentence $\neg A \neg x \wedge A(x \to \varphi)$ is equivalent to $@_x\varphi$.

The $@_x$ operator in some sense 'jumps', but note that it is not itself non-local, since the non-locality depends really on the values x can take. We will look in a moment at a more local kind of quantification. First, though let us draw the reader's attention to the expressivity of $@$.

In combination with the first-order quantifier, the $@_x$ modality is expressive. Indeed, in the case of relational models, it can express everything that an equivalent first-order language could express over the model's signature. The ***first-order formulae*** for N players over relational models for the propositions Ψ is given as follows, where $p \in \Psi$:

$$\varphi ::= p(x) \mid xR_iy \mid \neg\varphi \mid \varphi \wedge \varphi \mid \forall x\varphi.$$

A first-order assignment δ is a function assigning a single state to each first-order variable, the set of which we also write VAR, i.e. $\delta : \text{VAR} \to W$. Then we write $\mathcal{M} \vDash_\delta \varphi$ to mean that the assignment δ makes the first-order formula *true*, and define the relation as follows:

$$\begin{array}{lll}
\mathcal{M} \vDash_\delta p(x) & \text{iff} & \delta(x) \in [\![p]\!]_\mathcal{M} \\
\mathcal{M} \vDash_\delta xR_iy & \text{iff} & \delta(y) \in \bigcap \mathcal{N}_i(\delta(x)) \\
\mathcal{M} \vDash_\delta \neg\varphi & \text{iff} & \mathcal{M} \nvDash_\delta \varphi \\
\mathcal{M} \vDash_\delta \varphi \wedge \psi & \text{iff} & \mathcal{M} \vDash_\delta \varphi\ \&\ \mathcal{M} \vDash_\delta \psi \\
\mathcal{M} \vDash_\delta \forall x\varphi & \text{iff} & \text{for all } \delta' \sim_{-x} \delta,\ \mathcal{M} \vDash_{\delta'} \varphi.
\end{array}$$

A variable x is called '*free*' in a formula if it does not occur there under the scope of a quantification $\forall x$. The ***first-order language*** is the set of first-order formulae that have *one* free variable, and we overload notation and write $[\![\varphi(x)]\!]_\mathcal{M}$ to refer to the set of states that make $\varphi(x)$ true, in the following sense:

$$[\![\varphi(x)]\!]_\mathcal{M} = \{u \in W \mid \delta(x) = u \Rightarrow \mathcal{M} \vDash_\delta \varphi(x)\}.$$

(Here and in general, where $\varphi(x)$ denotes a first-order formula, we mean that x is the unique free variable in the formula φ.)

Then the language $\mathcal{L}_{N,\forall,@}$ is equi-expressive with this language, in the sense that the following are equivalent for any event X in the model $\mathcal{M}$:

1. There is a first-order formula φ such that $[\![\varphi]\!]_{\mathcal{M}} = X$.
2. There is a formula $\varphi \in \mathcal{L}_{N,\forall,@}$ such that $[\![\varphi]\!]_{\mathcal{M}} = X$.

This is proved by showing that there is a straightforward translation between the languages. We do not give the details here; cf. [Cate, 2005].

The other kind of quantification over states that we consider is known as '*bounded quantification*'. The hybrid 'binder' $\downarrow$ is used to assign a 'name' to the current state, in order later to refer back to it. Logically speaking, it is a kind of quantifier. The new clauses, to transform a language $\mathcal{L}$ to a binder language $\mathcal{L}_{\downarrow}$, are the following:

$$\varphi ::= \ldots \mid x \mid \downarrow x.\varphi,$$

where $x \in \text{VAR}$. And again we relativise the semantics to an assignment function δ, but this time the semantics of the quantifier are as follows:

$$\mathcal{M}, u \vDash_{\delta} \downarrow x.\varphi \quad \text{iff} \quad \mathcal{M}, u \vDash_{\delta[x \mapsto \{u\}]} \varphi,$$

where the new assignment $\delta[x \mapsto \{u\}]$ is the unique assignment δ' such that $\delta' \sim_{-x} \delta$ and $\delta'(x) = \{u\}$. Notice that this kind of quantification is *local* in our sense. In fact, it can be characterised model-theoretically as the fragment of first-order logic that is preserved under generated sub-models: [Areces *et al.*, 1999] show that the binder quantifier is equivalent to first-order bounded quantification, shown by Feferman [Feferman, 1968] to be the fragment of the first-order language that is invariant under generated sub-models. However, if we add the universal modality, then we are again back at the full first-order language, so the following are equivalent:

1. There is a first-order formula φ such that $[\![\varphi]\!]_{\mathcal{M}} = X$.
2. There is a formula $\varphi \in \mathcal{L}_{N,\downarrow,A}$ such that $[\![\varphi]\!]_{\mathcal{M}} = X$.

We can use any modal language with the binder, optimality operators and converse modality in order to define rationality. That is what is stated by Fact 2.5.

Fact 2.5. *If $\mathcal{L} \supseteq \mathcal{L}^{\emptyset}_{N,O,\downarrow,-1}$, then there is a formula in $\mathcal{L}$ that defines that player i is rational on relational models.*

Proof. $[\![\downarrow x.\, \bigcirc_i (\Diamond_i^{-1} x)]\!]$ is the event that the player i is rational. (This is immediate given Fact 2.4.) ■

Corollary 2.1. *If $\mathcal{L} \supseteq \mathcal{L}^{\emptyset}_{N,O,\downarrow}$, then there is a formula in $\mathcal{L}$ that defines that player i is rational on partitional models.*

An arguably more natural way to define rationality is to use the fact that the language can express what strategies players are playing. This means that the definition of rationality will be parametrised by the particular game being analysed, but that is not a problem since we will nonetheless give a generic formulation of rationality in terms of disjunctions over the strategy sets. In contrast to the previous definitions, this will work only in the presence of the constraints that hold on optimality operators, so it does not work for arbitrary monotonic operators $\bigcirc_i$, but since rationality only makes sense with respect to optimality operators that is not a problem. The final concern is about *cardinality*: if the cardinality of the strategy sets exceeds the cardinality of the language then we will *not* be able to define rationality in this way, and so would have recourse to one of the previous definitions.

The idea is very straightforward, and mirrors the generic definition of rationality in terms of second-order quantification. It relies essentially on us placing some some constraints on optimality operators, specifically on constraint 2 that we discussed after Proposition 2.1. That constraint says that optimality of a player's choice with respect to a formula depends only on the restriction defined by that formula. Therefore we do not need to quantify over *all events*, but only over those events that define restrictions of the game.[7] If the language is expressive enough, in terms of its basic propositions and its cardinality, then we can define each of those restrictions S with a formula φ_S. Finally, if the cardinality of the language suffices then we can use these φ_S's in order to 'simulate' the universal quantification with a conjunction.

Given some restriction $S \subseteq T$, of cardinality less than the cardinality of the language, then if either $T \subseteq \Psi$ or $\bigcup T_i \subseteq \Psi$, we can define the event that players play according to S, simply by taking an appropriate disjunction. So in the case where $T \subseteq \Psi$, we simply set

$$\varphi_S = \bigvee_{s \in S} \mathbf{s}.$$

Then, in case the cardinality of the language allows it, the following sentence would be a formula of the language:

$$\bigwedge_{S \subseteq T} (\Box_i \varphi_S \rightarrow \bigcirc_i \varphi_S).$$

When does the cardinality of the language allow it? The 'length' of this sentence is bounded by $2.\#(T) \times 2^{\#(T)}$. Therefore if the cardinality of the language is greater than this then the relevant sentence is a formula of the language. So if the language has proposition letters for strategies or for strategy profiles, and if κ, the cardinality of the language, is inaccessible and $\#(T) < \kappa$, then the sentence is in the language.

Furthermore, as Fact 2.6 states, on *monotonic* models the sentence will define rationality.

[7] Recall that an event is an arbitrary subset of the state space, whereas a restriction is a set of strategy profiles.

Fact 2.6. *On* monotonic *models,* $[\![\bigwedge_{S\subseteq T}(\Box_i\varphi_S \to \bigcirc_i\varphi_S)]\!]$ *defines the event that player* i *is rational.*

Proof. We show the direction where monotonicity is required: that the sentence being true entails that the player is rational. So suppose that

$$\mathcal{M}_G, u \vDash \bigwedge_{S\subseteq T}(\Box_i\varphi_S \to \bigcirc_i\varphi_S).$$

Then take any $X \in \mathcal{N}_i(u)$. We must show that $\xi_i(u) \in O_i(\xi(X))$. By the monotonicity of $\mathcal{N}_i(u)$ and the fact that $\xi^{-1}(\xi(X)) \supseteq X$, we have that $\xi^{-1}(\xi(X)) \in \mathcal{N}_i(u)$. In which case $u \vDash \Box_i\varphi_{\xi(X)}$, so by supposition $u \vDash \bigcirc_i\varphi_{\xi(X)}$, i.e. $\xi_i(u) \in O_i(\xi(X))$. ■

Notice that we have seen different kinds of definition of rationality. In the basic modal language (including the infinitary modal language), we had to talk specifically about the game in the language, so this was in some sense not a 'emphuniform' definition, in that it is parametrised by the game. The abstract definition using a converse modality and the hybrid binder, and the more explicit definition using second-order quantification, were on the other hand both 'uniform'. We do not have more to say about this, and it is not yet clear how to make precise the difference between uniform and non-uniform definition, and so how to establish that the less expressive modal languages might have a non-uniform, but no uniform, definition of rationality.

Another notion that is definable via a second-order approach or by infinite conjunctions is that of *common belief*. As long as the cardinality of the language is sufficient, the finitary basic modal language can define *mutual belief*, as follows:

$$\Box\varphi := \bigwedge_{i\in N}\Box_i\varphi.$$

The language described by Aumann [1999] does not include common belief in the language. Rather, it is only described there at the more informal semantic level as the (countable) intersection of iterated belief operators, so in our terminology it is $\Box^*$. (Recall that on relational models the countable intersection is equivalent to the arbitrary intersection.)

The other characterisation of common belief is in terms of the existence of an *evident event* (Fact 1.5 from Section 1.3). Recall that an 'evident' event is an event E such that at every state in E, every player believes he or she is in E, i.e. such that $\forall i \in N,\ R_i(E) \subseteq E$. In any language including $\mathcal{L}_{N,A,\overline{\forall}}$, we can express the fact that an event is evident: where $\delta(X) = E$, the following is equivalent to E being evident:

$$\mathcal{M}, u \vDash_\delta A(X \to \Box X).$$

So the these languages can certainly define common belief:

Fact 2.7. *Any language that contains* $\mathcal{L}_{N,A,\overline{\forall}}$ *can express common belief.*

Proof. The following formula defines common belief of φ, and is clearly in $\mathcal{L}_{N,A,\overline{\forall}}$.

$$\overline{\exists} X.(X \wedge A(X \rightarrow (\Box\varphi \wedge \Box X))).$$

■

However, we do not need to go as far as this in order to define common belief. One option would be to add the common belief operator to the language, and to ensure via appropriate axioms that it behaves as a common belief operator, i.e. to show that it is axiom-definable. That is a common method for including a common belief operator in a language, but is not something we pursue further here. Cf. [Lismont and Mongin, 1994; Fagin *et al.*, 1995; Heifetz, 1996].

Another extension to the language that we will consider is adding *fixpoint quantifiers*, which as is known can also be used to define common belief. After all, the *evident event* whose existence is asserted by the second-order formulation of quantified belief in Fact 2.7 just *is* the fixpoint for a certain operation involving mutual belief (cf. Fact 1.5). That is *why* fixpoint quantifiers, as we will see in a moment (Fact 2.8 below), are able to express common belief.

Fixpoint quantifiers are perhaps initially less natural to interpret than the other quantifiers, but they provide natural ways to express *iterative* concepts like the two key notions from the previous Chapter: common belief and the iterated elimination of non-optimal strategies. To extend a language $\mathcal{L}$ to a fixpoint language $\mathcal{L}_\nu$, we add this clause:

$$\varphi ::= \ldots \mid X \mid \nu X.\varphi,$$

where again X is a variable interpreted, by an assignment, as an event. However, we also add the following two restrictions:

1. φ is *positive in* X, meaning that X occurs under the scope of an even number of negation symbols $\neg$.

2. All operators occurring in φ are *monotonic*.

These two restrictions together mean that the function defined by $\varphi(X)$ is monotonic, i.e. that, in a similar way as with the case of non-strict dominance discussed above, where $\varphi(X)$ is a formula with X the only free variable, the following map is monotonic:

$$\begin{array}{rccc} F_\varphi: & 2^W & \rightarrow & 2^W \\ & E & \mapsto & [\![\varphi]\!]^{\{(X,E)\}}. \end{array}$$

That means, as we mentioned in the previous Chapter, that as a corollary of [Tarski, 1955, Theorem 1], there are both a largest and a smallest fixpoint for the function. The semantic clause for the fixpoint operator is the following:

$$\mathcal{M}, u \vDash_\delta \nu X.\varphi \quad \text{iff} \quad \exists E \subseteq |\mathcal{M}| : u \in E \subseteq [\![\varphi]\!]^{\delta[X \mapsto E]}_{\mathcal{M}}$$

This means that $\nu X.\varphi(X)$ is interpreted by the event that is the largest fixpoint for the map defined by $\varphi(X)$. Which entails, given Fact 1.5, that common belief is definable. (Actually we are using the generalisation of Fact 1.5 to monotonic neighbourhood models given in [Heifetz, 1996, Proposition 2.1].)

Fact 2.8. *The following are equivalent:*

1. $u \in \Box^{\infty}[\![\varphi]\!]$.
2. $u \vDash \nu X.(\Box\varphi \wedge \Box X)$.

Recall that on relational and topological models, $[\![\Box(\varphi \wedge \psi)]\!] = [\![\Box\varphi \wedge \Box\psi]\!]$. Then Corollary 2.2 is immediate:

Corollary 2.2. *On relational and topological models, the following are equivalent:*

1. $u \in \Box^{\infty}[\![\varphi]\!]$.
2. $u \vDash \nu X.\, \Box\, (\varphi \wedge X)$.

Corollary 2.2 is tight though, in the sense that it does *not* in general hold for neighbourhood models (monotonic or otherwise). On monotonic neighbourhood models this second fixpoint that we have defined $\nu X.\, \Box\, (\varphi \wedge X)$ is related to *co*-mutual belief (Definition 1.12), but this is not something we investigate further here.

The trick to proving Theorem 1.1 in a relatively simple manner was to observe that not only is common belief a kind of fixpoint, but so is the outcome of the iterated elimination of non-optimal strategies, when the optimality operator is monotonic. Indeed, in that case, as Fact 2.9 states, we can define the outcome of elimination of non-optimal strategies.

Fact 2.9. $\xi([\![\nu X.\, \bigcirc\, X]\!]) = O^{\infty}$.

The last method we consider for defining common belief is perhaps less elegant, but certainly does the job in a clear way. That is to define it directly as the infinite conjunction that it is always defined as intuitively: 'everybody believes that φ, everybody believes that everybody believes that φ,...'. As we have seen, in non-relational neighbourhood models it is possible to separate finitary and transfinite common belief. Cardinality considerations show us that in no $\mathcal{L}_{N,\kappa}$ is there a sentence that expresses '*true*' common belief in arbitrary neighbourhood models.

However, any properly infinitary version of the basic modal language can define *finitary* common belief:

Fact 2.10. *Any language including $\mathcal{L}_{N,\kappa}$, where $\kappa > \aleph_0$, can define finitary common belief $\Box^{*}\varphi$.*

Proof. The following formula defines common belief in φ, and is in $\mathcal{L}_{N,\kappa}$:

$$\bigwedge_{m\in\mathbb{N}} \underbrace{\Box\ldots\Box}_{m \text{ times}}\varphi.$$

■

For 'bigger' versions of common belief, note that if the cardinality of the language is κ, then it can define α-order mutual belief for any ordinal $\alpha < \kappa$. But we should re-iterate that in the relational models that are used in the game theory literature, any suitably infinitary modal language can define common belief.

Suppose that we were to have formulae $\mathbf{r}$ in our language, to be interpreted as: 'player i is rational'. We have seen that this can be definable, in a suitably expressive language. It could also be primitive. Either way, certain axioms will be valid with respect to this formula. (Recall that validity means being true in every model.) Most notably, an axiom similar to that considered by de Bruin [2004] will be valid:

$$\mathbf{r_i} \rightarrow (\Box_i\varphi \rightarrow \bigcirc_i\varphi) \qquad \mathbf{r_i}Dis$$

Then if we write $\mathbf{r}$ for $\bigwedge_{i\in N}\mathbf{r}_i$, and similarly for $\Box$ and $\bigcirc$, then clearly the following statement is derivable from the axioms $\mathbf{r}_iDis$:

$$\mathbf{r} \rightarrow (\Box\varphi \rightarrow \bigcirc\varphi) \qquad \mathbf{r}Dis$$

Furthermore, the following axiom is valid for the fixpoint operator:

$$\nu X.\varphi \rightarrow \varphi[X \mapsto \nu X.\varphi] \qquad \nu Dis,$$

where $\varphi[X \mapsto \psi]$ denotes the formula obtained by replacing all occurrences of x in φ by ψ. That axiom along with the following *proof rule* were introduced in [Kozen, 1983].

$$\frac{\psi \rightarrow \varphi[x \mapsto \psi]}{\psi \rightarrow \nu x.\varphi}\ \nu Ind$$

It is straightforward to show that the proof rule is 'sound', meaning that when its premise is valid then so is its conclusion. Furthermore, *in the case of relational models*, the axiom and proof rule have been shown by Walukiewicz [1995] to generate, along with standard axioms and rules for the language $\mathcal{L}_N$, the $\mathcal{L}_{N,\nu}$-logic of those models. I.e. he shows that there is a complete axiomatisation of the logic. The $\mathcal{L}_{N,\nu}$-logic of *neighbourhood* models does not yet have a complete axiomatisation, nor do we have a complete axiomatisation of the $\mathbf{r}_i$ operator. It is beyond the scope of this Chapter to axiomatise any complete logics, but we will now show that $\mathbf{r}Dis$, νDis and νInd are sufficient to give a proof of Theorem 1.1.

Theorem 1.1 can be stated syntactically as follows:

(Th1.1) $(\Box^\infty\mathbf{r} \wedge \mathbf{r}) \rightarrow \nu X.\bigcirc X$,

where $\Box^\infty\varphi$ is an abbreviation of $\nu X.\ \Box\ (\varphi \wedge X)$, cf. Corollary 2.2.

Fact 2.11. *(Th1.1) is true iff rationality and common belief of rationality imply that the players will not play strategies that survive the elimination of non-optimal strategies. (This uses Fact 2.9.)*

By Fact 2.11, the validity of (Th1.1) on relational models *is equivalent to* Theorem 1.1.

Theorem 2.1. *(Th1.1) is valid.*

Proof. The following formula is an instance of the axiom $\mathbf{r}Dis$
(set $\varphi := \Box^\infty\mathbf{r} \wedge \mathbf{r}$):

$$\mathbf{r} \to (\Box(\Box^\infty\mathbf{r} \wedge \mathbf{r}) \to \bigcirc(\Box^\infty\mathbf{r} \wedge \mathbf{r})),$$

Given that $\Box^\infty\mathbf{r}$ is really $\nu X.\ \Box\ (\mathbf{r} \wedge X)$, the following is equivalent to an instance of νDis (setting $\varphi := \Box(X \wedge \mathbf{r})$):

$$\Box^\infty\mathbf{r} \to \Box(\Box^\infty\mathbf{r} \wedge \mathbf{r})$$

Putting these two together via some simple propositional logic, we obtain:

$$(\Box^\infty\mathbf{r} \wedge \mathbf{r}) \to \bigcirc(\Box^\infty\mathbf{r} \wedge \mathbf{r}).$$

This last formula is of the right shape to apply the rule νInd
(with $\psi := \Box^\infty\mathbf{r} \wedge \mathbf{r}$ and $\varphi := \bigcirc X$), to obtain:

$$(\Box^\infty\mathbf{r} \wedge \mathbf{r}) \to \nu X.\ \bigcirc\ X.$$

■

Notice that we say that Theorem 2.1 asserts that (Th1.1) is valid including on neighbourhood models. However, recall that Corollary 2.2, that makes it sensible to abbreviate $\Box^\infty\mathbf{r}$ as $\nu X.\ \Box\ (\mathbf{r} \wedge X)$, applies only to *relational* and to *topological* models.

We do not pursue questions of axiomatising the logics of any of these languages, i.e. of providing syntactic rules of manipulation that could be used to derive all the formulae of them that are valid.

2.3 Complete models

Rather than taking *states* as primitive, early work in the epistemics of game theory was based on the idea of so-called 'types' [Harsanyi, 1968], and authors are still divided in their approaches, so 'type space' models are often studied, rather than state space models. A *type* for a player can be thought of as the way a player might be, so it is a '*possible player*', where states are '*possible worlds*'. A type for a player specifies

the beliefs of that player, defined in terms of the other players, and also specifies what strategy that player will play.[8] We will show in this Section how to interpret the different languages we considered on type-space models, and translate between type-space and state-space models. There is already a sort of common understanding that the two ways of modelling are in some sense equivalent, but here we map out exactly in what sense they are equivalent. We then turn to consider a kind of 'fullness' property of models, 'assumption-completeness'. We show what this means on state-space models, and then turn back to look at it in the way considered in [Brandenburger and Keisler, 2006], where it is studied seriously for the first time, in order to prove a positive result that there are, in the sense of that paper, assumption-complete models for the basic modal language and some infinitary versions of it.

Definition 2.5. A ***type-space model*** for the game $(T_i, >_i)_{i \in N}$ is a structure

$$(W_i, R_i, \xi_i)_{i \in N},$$

where W_i is a set, called i's '*states*', or '*types*', $\xi_i : W_i \to T_i$, and $R_i \subseteq W_i \times W_{-i}$.

Here we are only considering *relational* type-space models; as far as we are aware neighbourhood or topological type-space models have not been considered in the literature. As with the state-space models that we have so far considered, R_i gives i's *information*, which here is taken to mean i's information concerning the other players. Indeed, as we shall see in a moment, there is no way to represent a player's uncertainty about her *own* type in these models, so players are positively and negatively introspective.

On these many-sorted models, there are two natural ways for us to make one single-sorted domain. The first is to take the *union* of the W_i's; the other is to take the *product*. We favour the latter approach, mainly because it makes the connection between type-space and state-space models easier. However, the former approach is taken in [Brandenburger and Keisler, 2006], in the special case of 2-player type-space models, and so we will return to it below.

For the product case, the easiest way to see how to interpret languages is to define a state-space model that is, intuitively, equivalent to the type-space model, and then interpret the language on that.

Definition 2.6. Given a type-space model $\mathcal{M} = (W_i, R_i, \xi_i)_{i \in N}$, we define the state-space model

$$\mathcal{S}(\mathcal{M}) = (W = \prod_{i \in N} W_i, R'_i, \xi')_{i \in N},$$

where

$$R'_i(u) = \{v \in W \mid u_i = v_i \text{ and } v_{-i} \in R_i(u_i)\},$$

[8] There are other formulations of type in which a type is not taken to specify a strategy, but that difference is not of any conceptual or mathematical significance, since it is easy to translate between the two formulations.

and

$$\xi(u) = (\xi_i(u_i))_{i \in N}.$$

Now, it is not possible to give a formal statement that these models are equivalent, but it should be intuitively clear that we have captured the type-space model in this state-space model.

What about the reverse direction? That is, given some state-space model, can we write down a type-space model that is equivalent to it? The problem here is that because in type-space models players are in effect certain (and correct) about their own type, they are fully introspective.

Fact 2.12. *In $\mathcal{S}(\mathcal{M})$, all players are positively and negatively introspective.*

Proof. Let $\mathcal{S} = (\prod_{i \in N} W_i, R_i, \xi)_{i \in N}$. First of all notice that for any states $u, v \in W = \prod_{i \in N} W_i$, we have the following Lemma, whose proof is immediate from Definition 2.6.

Lemma 2.1. *We have the following entailments:*

1. *If $v \in R_i(u)$ then $u_i = v_i$.*
2. *$R_i(u) = R_i(v)$* iff *$u_i = v_i$.*

Then recall that positive introspection is the following property (see Definition 1.6.4$^{\text{r}}$)

$$R_i(R_i(u)) \subseteq R_i(u).$$

So take any $w \in R_i(R_i(u))$. Then there exists $v \in R_i(u)$ such that $w \in R_i(v)$. Since $v \in R_i(u)$, then by Lemma 2.1.1., $v_i = u_i$, so by the Lemma 2.1.2, $R_i(v) = R_i(u)$. But recall that $w \in R_i(v)$; so $w \in R_i(u)$ as required for positive introspection.

Negative introspection (Definition 1.6.5$^{\text{r}}$) is the following property:

$$\{v, w\} \subseteq R_i(u) \Rightarrow w \in R_i(v).$$

So take any $\{v, w\} \subseteq R_i(u)$. By Lemma 2.1.1, $v_i = u_i = w_i$, so by Lemma 2.1.2, $w \in R_i(v)$ as required. ■

This means of course that it is not possible to find a type-space model that faithfully represents *every* state-space model, since there are state-space models that are not positively and negatively introspective. However, we will now describe a translation that does faithfully represent every positively and negatively introspective relational model as a (relational) type-space model.

So let us take a relational state-space model $\mathcal{S} = (W, R_i, \xi)_{i \in N}$.

Definition 2.7. First, we define an equivalence relation $\sim_i$ on the state space:

$$u \sim_i v \text{ iff } R_i(u) = R_i(v) \,\&\, \xi_i(u) = \xi_i(v).$$

The relation $\sim_i$ in Definition 2.7 is supposed to capture the identity of types: $u \sim_i v$ means that i's 'type' is the same at u and at v. It is not actually necessary to use this for the definition, but we use it to ensure that if we go back and forth, from a type model to a state model and back again, we do not add any new types.

So we use this equivalence relation $\sim_i$ in order to give the types in the type-space model $\mathcal{T}(\mathcal{S})$ that we will define. Given the equivalence relation $\sim_i$, we write $[u]_i$ to mean the *equivalence class* of u, i.e.

$$[u]_i = \{v \in W \mid u \sim_i v\}.$$

This allows us to define the translation $\mathcal{T}(\_)$.

Definition 2.8. Let $\mathcal{T}(\mathcal{S}) = (\{[u]_i \mid u \in W\}, R'_i, \xi'_i)_{i \in N}$, where

$$[s]_i R'_i \prod_{j \neq i} [u_j]_j \text{ iff } \exists v \in R_i(s) : \forall j \in N - \{i\},\ v_j \in [u_j]_j.$$

and

$$\xi'_i([u]_i) = (\xi(u))_i.$$

The following Proposition states that the translation makes sense, in that if we start with a type-space model, translate it to a state-space model and then translate the resulting state-space model to a type-space model, that second translation is in effect a 'translating back', so that we end up with something isomorphic to the original model:

Proposition 2.3. *For any type-space model $\mathcal{M}$, we have $\mathcal{T}(\mathcal{S}(\mathcal{M}))$ isomorphic with $\mathcal{M}$.*

We have formalised, then, the intuitive connection between state-space models, a mainstay of formal epistemology, and type-space models.

A recurring issue in the literature on epistemic analysis of games involves defining a space of *all possible* beliefs of the players and whether such a space exists. We will look now in detail at one way of cashing out the notion that a model is a space of all possible beliefs. Specifically, we consider the property of *assumption-completeness*, first introduced in [Brandenburger, 2003].

Following [Brandenburger and Keisler, 2006], we use the word 'assumption' to refer to the content of a player's information.

Definition 2.9. If in some type-space model, $R_i(u) = [\![\varphi]\!]$, i.e. the player's *information* is determined by $[\![\varphi]\!]$, then we say that φ is i's ***assumption***, or equivalently that i *assumes* φ, at u.

This notion of assumption is closely related to the only-knowing operator studied by Levesque [Levesque, 1990] (cf. [Halpern and Lakemeyer, 2001]). A player's *assumption* is her strongest belief: the conjunction of all her beliefs (equivalently, a belief that implies all her other beliefs). This definition of "assumption" might seem

strange, and we certainly do not claim that it captures the common-sense meaning of the English word "assumption".

Notice that in neighbourhood models, unless the model is relational (i.e. monotonic and contains its core, cf. Definition 1.17), there is no obvious analogue of a player's assumption at a given state, since the information is given not by a single formula but by a collection of them. We therefore will only consider relational models for the rest of this Section.

The property of assumption-completeness asserts that *every definable set of i's opponents' types* corresponds to a belief somewhere of i. That is, every possible *definable* configuration of i's beliefs is represented somewhere in the model. Since it is a property that talks about definability, it is relative to a language.

Definition 2.10. The type-space model $\mathcal{T} = (W_i, R_i, \xi)$ is ***assumption-complete*** for $\mathcal{L}$ just if, for all players i:

$$\forall \varphi \in \mathcal{L}, \exists u_i \in W_i : R_i(u_i) = (\llbracket \varphi \rrbracket_{\mathcal{T}}) \cap W_{-i}.$$

Let us say that a type-space model is "*non-trivial*" when, for every player i, there is some type t_i that 'rules out' some type of some other player $j \neq i$, i.e. such that $(R_i(t_i))_j \neq T_j$.

Definition 2.11. Assumption-completeness as a property of a language just states that there exists *some* non-trivial model that is assumption-complete for that language. Conversely, if there is *no* non-trivial assumption-complete model for $\mathcal{L}$, then we say that $\mathcal{L}$ is ***assumption-incomplete***.

We postpone discussing the significance or otherwise of assumption-completeness until later in this Section; for now we turn attention to finding an analogue of it on state-space models.

How are we to define assumption-completeness for state-space models? The naïve approach, based on a simplistic reading of [Brandenburger and Keisler, 2006], would be to say that a state-space model $\mathcal{S}$ is assumption-complete for $\mathcal{L}$ just if for any $\varphi \in \mathcal{L}$, there is $u \in W$ such that $R_i(u) = \llbracket \varphi \rrbracket_{\mathcal{S}}$. Let us call this, for the purposes of our discussion, the 'tentative' definition of assumption-completeness. However, it is not difficult to see that this is not an innocent approach. For then even very simple languages are not assumption-complete:

Fact 2.13. *For the tentative definition of assumption-completeness just proposed, no language with disjunction has an assumption-complete introspective relational model, in which the language can define more than one event.*

Proof. This is easy to see: Take any relational model $\mathcal{M}$ satisfying positive and negative introspection, and two formulae φ and ψ of the language such that (in $\mathcal{M}$) $\llbracket \varphi \rrbracket \neq \llbracket \psi \rrbracket$, $\llbracket \varphi \rrbracket \not\subset \llbracket \psi \rrbracket \not\subset \llbracket \varphi \rrbracket$. Suppose towards a contradiction that $\mathcal{M}$ is assumption-complete according to the tentative definition. Then there would be a state u such that

$R_i(u) = [\![\varphi]\!]$. Since the language is closed for disjunction, there is also a state w such that $R_i(w) = [\![\varphi]\!] \cup [\![\psi]\!]$. Then take any $x \in [\![\varphi]\!]$ and any $y \in [\![\psi]\!] - [\![\varphi]\!]$. Now since wR_ix and wR_iy, then we know that xR_iy. But then since uR_ix, uR_iy, which is a contradiction. So the tentative definition will not do. ■

Since type-space models are fully introspective, obviously the proposed tentative definition is not acceptable. We want a more meaningful definition of assumption-completeness of course, one that is faithful to the definition on type-space models.

Assumption-completeness is explored in [Brandenburger and Keisler, 2006], where a number of results are established, and connections to modal logic are mentioned. As in that paper, we will henceforth restrict our attention to ***two-player*** models, fixing $N = \{a, b\}$. There may well be further issues to be discovered in the consideration of models with more players N, but the essential character of the issue is captured in the two-player case. We will now give a proper definition of assumption-completeness for models.

Definition 2.12. A model $\mathcal{S} = (W, R_a, R_b, \xi)$ is ***assumption-complete*** for a language $\mathcal{L}$ just if for any $\varphi \in \mathcal{L}$, for $\{i, j\} = \{a, b\}$, there exists $y \in W$ such that the following two conditions hold:

- $\forall x \in [\![\varphi]\!], \exists v \in R_i(y) : R_j(v) = R_j(x)$ & $\xi_j(v) = \xi_j(x)$;
- $\forall v \in R_i(y), \exists x \in [\![\varphi]\!] : R_j(v) = R_j(x)$ & $\xi_j(v) = \xi_j(x)$.

We say that $\mathcal{S}$ is assumption-complete *tout court* when it is assumption-complete for a and for b.

Definition 2.12 might seem more long-winded than Definition 2.10, but it is equivalent to the definition for type-space models, in the following sense:

Theorem 2.2. *The following equivalences hold for any type-space model $\mathcal{T}$ and any state-space model $\mathcal{M}$:*

- *$\mathcal{T}$ is assumption-complete (in the sense of Definition 2.10) iff $\mathcal{S}(\mathcal{T})$ is assumption-complete (in the sense of Definition 2.12).*
- *$\mathcal{M}$ is assumption-complete iff $\mathcal{T}(\mathcal{M})$ is assumption-complete.*

Thus we have found the "correct" definition of assumption-completeness for single-sorted models.

The central result in [Brandenburger and Keisler, 2006] is an impossibility result: that the first-order language is assumption-incomplete.

The result that Brandenburger and Keisler prove is, in effect

Theorem 2.3 ([Brandenburger and Keisler, 2006, Theorem 5.4])**.** *There are* no *assumption-complete models for any language extending $\mathcal{L}_{N,\forall,@}$.*

Although we refer to [Brandenburger and Keisler, 2006] for the proof[9], let us say that it revolves around proving that a formula expressing the following statement, or some sub-formula of it, must be unsatisfiable on any model:

(BK) Ann believes that Bob's assumption is that Ann believes that Bob's assumption is wrong.

We will now look at how to formalise (BK) in some language for *state-space models*. We will see that this sentence (BK) is a sort of many-player, so *interactive* version of Russell's paradox. We therefore recommend it for further study also outside the field of game theory. For our present purposes, the sentence (BK) is also crucial to the proof of Theorem 2.4, so we investigate it in some detail in what follows. We will show that the (*implicit*) KD45 nature of type-space models is crucial to the particular argument used in [Brandenburger and Keisler, 2006] to prove Theorem 2.3. In order to do that, we first need to show how to formalise (BK) in $\mathcal{L}_{\{a,b\},\downarrow,A}$, so for the time being *we return again to state-space models* with two players a (Ann) and b (Bob).

The sentence 'Ann believes that Bob's assumption is wrong' can be formalised as follows:

$$\Box_a \downarrow x.\, \Box_b \neg x$$

And 'Bob assumes that φ' can be written as

$$\downarrow x.A(\varphi \equiv \downarrow y.(A(x \rightarrow \Diamond_b y)))$$

Putting these two together, we get that the formal translation of (BK) can be written as:

$$\Box_a \downarrow x.A(\Box_a \downarrow z.\, \Box_b \neg z \equiv \downarrow y.(A(x \rightarrow \Diamond_b y)))$$

We will use the following Facts to reveal some hidden premises in the informal version of Brandenburger and Keisler's argument. Note that this does not detract from the validity of Theorem 2.3. All that it does is to *suggest* that the implicit KD45 nature of type-space models, most notably the positive and negative introspection aspect (cf. Fact 2.12) might be crucial to its proof.

Fact 2.14. *The formal translation of the sentence (BK) is satisfiable on some state-space model, as long as we do not require all of D, 4 and 5 to hold.*

Proof. We give counterexamples in order to prove this Fact.

D. The first and simplest example is perhaps no surprise, since Brandenburger and Keisler explicitly add the condition D to their type-space models. The state-space model that we use as a counterexample, where D does not hold, is

$$\mathcal{M}_D = (\{u\}, R_a = \emptyset, R_b = \{(u,u)\}, \xi),$$

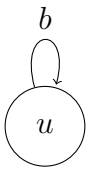


Figure 2.3: A model without D_a

where ξ is, as in the next two models, arbitrary. A diagram of the relational structure is given in Figure 2.3.

$\mathcal{M}_D$ is KD45 *except* that Ann lacks D. Furthermore, at u, the formal translation of (BK) holds.

4. In the model $\mathcal{M}_4$, we drop transitivity for a.

$$\begin{aligned}\mathcal{M}_4 = \quad & (\{u,v,w\}, \quad R_a = \{(u,v),(v,v),(v,w),(w,w)\},\\ & \qquad\qquad R_b = \{(u,u),(v,u),(w,w)\},\xi).\end{aligned}$$

Figure 2.4 represents $\mathcal{M}_4$.

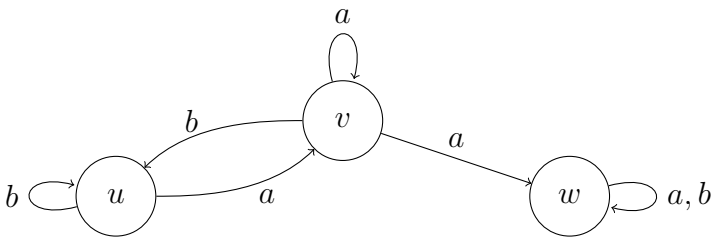


Figure 2.4: A model in which Ann lacks positive introspection

$\mathcal{M}_4$ has all introspection properties except that Ann lacks 4. Furthermore, at u, (BK) is true.

5. In this model $\mathcal{M}_5$, Ann lacks negative introspection. (In fact she lacks confidence in her own beliefs.)

$$\begin{aligned}\mathcal{M}_5 = \quad & (\{u,v,w,x\}, R_a = \{(u,v),(u,w),(u,x),(v,w),(w,x),(x,x)\},\\ & R_b = \{(u,v),(v,v),(w,v),(x,w)\}\end{aligned}$$

$\mathcal{M}_5$ has all introspection properties except that Ann lacks 5 (and indeed mT). Furthermore, at w, (BK) holds.

[9] A syntactic proof of [Brandenburger and Keisler, 2006, Theorem 5.4], is found in [Pacuit, 2007]. Though note that the language used in [Pacuit, 2007] is $\mathcal{L}^{\Psi}_{N,\downarrow,A}$, which as we have noted is a syntactic variant of first-order logic.

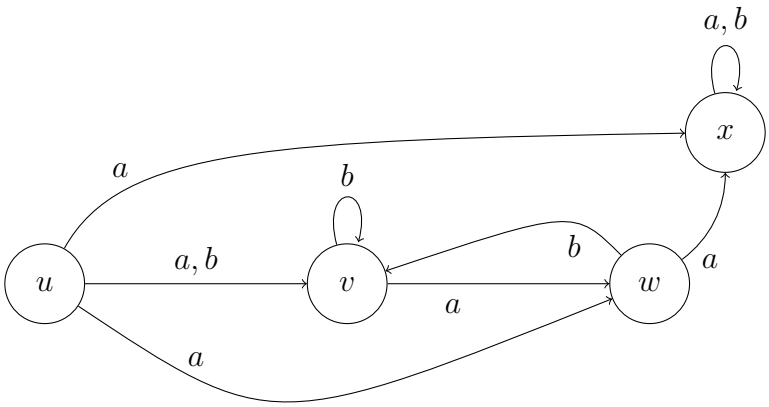


Figure 2.5: A model in which Ann is not negatively introspective.

■

That is, if introspection fails then the sentence (BK) is consistent. It might seem puzzling that an informal argument is given in [Brandenburger and Keisler, 2006, Section 1] to the effect that (BK) is not satisfiable, an argument where the word "introspection" is never used, nor is any concept like it employed. Fact 2.14 illustrates that corners were cut in the informal argument. Let us spend a moment to unpick the threads of the argument.

Suppose towards a contradiction that (BK) is true. We quote the following argument:

> "*To get the impossibility, ask: Does Ann believe that Bob's assumption is wrong? If so, then in Ann's view, Bob's assumption, namely 'Ann believes that Bob's assumption is wrong', is right. But then Ann does not believe that Bob's assumption is wrong, which contradicts our starting supposition. This leaves the other possibility, that Ann does not believe that Bob's assumption is wrong. If this is so, then in Ann's view, Bob's assumption, namely 'Ann believes that Bob's assumption is wrong', is wrong. But then Ann does believe that Bob's assumption is wrong, so we again get a contradiction.*" – op.cit.

So (BK) must be false. Yet we have illustrated models in which, given the failure of D, 4 or 5, (BK) *is* true. So let us look one-by-one at where each is implicitly used in the argument.

If Ann is allowed to be inconsistent in her beliefs, then of course she can believe what (BK) asserts that she believe! It is disarming to note that the quoted argument has this unseen auxiliary premise. Where in that argument is this auxiliary premise hidden? Notice that at u in the model in Figure 2.3, Ann *does* believe that Bob's assumption

is wrong. Call that proposition q. She also believes that Bob's assumption is q. She believes everything, in fact, because $\forall X, R_a(u) \subseteq X$! In particular, she believes that Bob's assumption is right. The additional step taken in the quoted argument is to infer that this means that she does not believe that Bob's assumption is wrong.

In the case of positive introspection it is more interesting to ask the question where does the quoted argument hide the auxiliary premise. Again, notice that this time $u \vDash \Box_a \downarrow x.\Box_b \neg x$; that is, Ann *does* believe that Bob's assumption is wrong. We call the proposition that Bob's assumption is correct 'p'. The argument quoted above drew a contradiction from this situation:

> "*[. . .] then in Ann's view, Bob's assumption, namely 'Ann believes that Bob's assumption is wrong', is right. But then Ann does not believe that Bob's assumption is wrong, which contradicts our starting supposition.*"

Ann *does* believe that she does not believe $\neg p$. That is: $\Box_a \neg \Box_a \neg p$. And we already have $\Box_a \neg p$. But we do not have $\Box_a \Box_a \neg p$, which would (along with Ann's consistency), yield a contradiction. For that, we need positive introspection.

In the case of negative introspection, all that is needed is the property mT mentioned in Section 1.3, that says that if a player believes that she believes something, then she believes it. That, recall, is entailed by 5. Again we compare this to the quoted argument. This time, $u \vDash \neg \Box_a \downarrow x. \Box_b \neg x$: Ann does not believe that Bob's assumption is wrong; she considers it possible that it is right.

> "*If this is so, then in Ann's view, Bob's assumption, namely 'Ann believes that Bob's assumption is wrong', is wrong. But then Ann does believe that Bob's assumption is wrong, so we again get a contradiction.*"

This goes a little quickly. "Ann believes that Bob's assumption is wrong", certainly, is wrong; but the movement to conclude that Ann believes that she does not believe that Bob's assumption is wrong involves using mT, $\Box_a(\Box_a \varphi \to \varphi)$. This says that Ann does not believe herself to have any false beliefs. In particular, it says that Ann believes that: if she believes that Bob is correct then Bob is correct. So in any possibility that Ann considers in which Bob is correct, she considers it possible that Bob is correct. Otherwise put: In any possibility that she considers in which Bob's assumption is right, she does not believe that Bob's assumption is wrong ($\Box_a(\neg p \to \neg \Box_a p)$). So any such possibility cannot be part of Bob's assumption, which is specified by (BK) to be precisely those possibilities at which Ann *does* believe that Bob's assumption is wrong. But if a possibility is not part of Bob's assumption, then at that possibility, Bob's assumption is incorrect. Therefore there can be no possibility for Ann at which Bob's assumption is correct. I.e. she believes that Bob's assumption is wrong.

So we have seen that assumption-completeness does depend on some properties of the players' beliefs, and shown how that fact plays out formally and informally. This suggests that the following statement may have been made too hastily:

> "*our impossibility phenomenon is not affected by the players' beliefs about themselves.*" [Brandenburger and Keisler, 2006].

Let us turn now to look at the question why we would be interested in assumption-completeness in the first place.

The idea presented in [Brandenburger and Keisler, 2006] that is supposed to underlie assumption-completeness is that the language for non-i states should be accessible to i. And if it is accessible to i then he should be able to *assume* (in this artificial sense) any member of it. So Theorem 2.3 was taken to be a *limiting* result, and one that should be of significance for game theory:

> "*[O]ur impossibility theorem says: If the analyst's tools are available to the players, there are statements that the players can think about but cannot assume. The model must be [assumption-]incomplete. This appears to be a kind of basic limitation in the analysis of games.*" – [Brandenburger and Keisler, 2006]

Further arguments are surely needed before we have any reason to accept that assumption-completeness is a necessary condition of the language being 'available' to the players in their reasoning about the model. Assumption-completeness is certainly technically interesting, and there may be some philosophical significance to it, but this notion of 'availability to the players' needs to be developed further, since it is not clear why the presence of a certain state in the model means that players do have the language 'available' to them, or that they 'can' assume something.

As Brandenburger and Keisler point out, the existence of assumption-complete models is not only of theoretical interest [Brandenburger and Keisler, 2006]. Epistemic conditions discussed in game theory sometimes involve notions of "completeness" of the underlying belief model, notions that are similar in flavour to assumption-completeness. These occur for example in two analyses: in Battigalli and Siniscalchi's analysis of extensive-form rationalisability [Battigalli and Siniscalchi, 2002] and Brandenburger, Friedenberg and Keisler's analysis of iterated admissibility [Brandenburger *et al.*, 2008].

Let us note also that it is not clear in what sense the putative *equivalence* that for example Theorems 1.1 and 1.2 were intended to establish is clearer in an assumption-complete belief model, where *all* hierarchies of beliefs are present. The only argument that might be given towards thinking that the equivalence might be clearer is simply that as long as rationality and common belief of rationality are expressible in the model, there will be a state in which each player assumes precisely that.

Nonetheless, let us grant that there might be some interest in finding languages that are assumption-complete. Given the above interpretation of Theorem 2.3, a natural question[10] is: can one define instead a restricted set of "tools" which *can* be "available" to the players, and which are also useful for the analyst? If Theorem 2.3 shows that the first-order language is too powerful a tool to be available, what about weaker languages?

[10]This is also raised in [Brandenburger and Keisler, 2006, Section 2].

We will look at one such case, showing that the infinitary modal language has assumption-complete models. In fact, we show a slightly stronger result, that it has "full" assumption-complete models, in the following sense (cf. Definition 1.4):

Definition 2.13. The type-space model $\mathcal{M} = (W_i, R_i, \xi_i)_{i \in N}$ is ***full*** for the game $G = (T_i, \geq_i)_{i \in N}$ just when

$$\forall i \in N, \forall s_i \in T_i, \exists u_i \in W_i : \xi_i(u_i) = s_i.$$

In order to bring things closer to the way they are presented in [Brandenburger and Keisler, 2006], we now slightly change our formal framework, in effect making a notational change. Rather than think of type-space models as state-space models by taking *products* of the type spaces, we now think of type-space models as *two-sorted* state-space models (recall that there are only two players), with a-states and b-states forming a partition of the state-space. That is, a language is now interpreted, as Brandenburger and Keisler propose, over the *union* of the domain of a- and b-states. Given a type-space model $(W_a, W_b, R_a, R_b, \xi_a, \xi_b)$, we consider the single-sorted model $(W_a \cup W_b, R_a \cup R_b, \xi_a \cup \xi_b, W_a)$. So this model is of the form (W, R, ξ, W_a). It is a model for a modal language with a single unary modality $\Box$ for the 'belief' relation, whose interpretation varies depending on whether we start in an Ann state (an element of W_a) or a Bob state (element of $W - W_a$). We will assume each language to have a nullary modality ♀ for W_a, i.e. to distinguish between Ann and Bob's domains. Now assumption-completeness of a model of this kind with respect to a language $\mathcal{L}$ is the property that for any $\varphi \in \mathcal{L}$ that defines a subset of W_a (resp. W_b), there is a $u_\varphi \in W_b$ (resp. W_a) such that $R(u_\varphi) = [\![\varphi]\!]$.

Note that this really is just a notational variant of the type-space model, but it will make proving the next Theorem more straightforward. So we interpret what are effectively 'one-player' versions of the languages we presented above on this model, the only addition being a proposition letter ♀ that is interpreted to mean the state is an Ann type ($[\![♀]\!] = W_a$). Since there are not several players, we now write $\mathcal{L}_\Diamond$ for the basic modal language. This might seem like an odd approach, but it is very close to the way things are presented by Brandenburger and Keisler.[11] It also allows us to prove directly the existence of a full assumption-complete model for the infinitary modal language:

Theorem 2.4. *For* any *game G, there are full assumption-complete models for $\mathcal{L}^{\Psi}_{\Diamond,O,\kappa}$ where for every strategy s_i there is a proposition letter $\mathbf{s_i} \in \Psi$.*

Proof. We define a semantic analogue to the 'canonical model' used in completeness proofs, cf. [Chellas, 1980], and show that it is assumption-complete.

Let us write $\mathcal{L}$ for $\mathcal{L}^{\Psi}_{\Diamond,O,\kappa}$. We also write $\Gamma \vDash \bot$ to mean that the set of sentences $\Gamma \subseteq \mathcal{L}$ is not jointly satisfiable, i.e. that there is *no* model $\mathcal{M}$ with a point u such that

[11] Actually, they consider a slightly different two-sorted quantification, writing $\exists x_a \varphi(x)$ to mean 'there is an Ann-type such that φ is true of it'. We would write this $\exists x.(@_x ♀ \wedge \varphi(x))$.

$\forall \gamma \in \Gamma, \mathcal{M}, u \vDash \gamma$. And define the maximally satisfiable sets $\textsc{mss}\ (\mathcal{L})$ as the following set:

$$\{\Gamma \subset \mathcal{L} \mid \Gamma \nvDash \bot \ \& \ \forall \Gamma' \supset \Gamma,\ \Gamma' \vDash \bot\}.$$

Finally, let the model $\mathcal{C}$ be $(\textsc{mss}\ (\mathcal{L}), R, O, \xi)$, where:

- $\Gamma R \Delta$ iff $\forall \delta \in \Delta, \Diamond\delta \in \Gamma$,
- $O(X) = \left\{\Gamma \in \textsc{mss}\ (\mathcal{L}) \middle| \exists \bigcirc \varphi \in \Gamma : \{\Delta \in \textsc{mss}\ (\mathcal{L}) \mid \varphi \in \Delta\} \subseteq X\right\}$,
- $\xi_i(\Gamma)$ is the $s_i \in T_i$ such that $\mathbf{s_i} \in \Gamma$.

It is possible to show by induction that in this model the is an exact match between the syntactic and the semantic structures, a fact, whose proof is standard, often referred to as a Truth Lemma:

Lemma (Truth Lemma)**.** $\mathcal{C}, \Gamma \vDash \varphi \Leftrightarrow \varphi \in \Gamma$.

Now take any definable subset X of $W^a = \{\Gamma \in \textsc{mss}\ (\mathcal{L}) \mid \Gamma \vDash \text{♀}\}$. (This is meant to be without loss of generality; for the other case, replace here and in what follows a with b and ♀ with $\neg$♀.) Then there is some formula $\varphi \in \mathcal{L}$ such that $[\![\varphi]\!]_{\mathcal{C}} = X$. We know that φ is equivalent to $\varphi \wedge$ ♀ (because otherwise $\varphi \wedge \neg$♀ would be satisfiable, in which case there would be some $\Delta \in [\![\varphi \wedge \neg\text{♀}]\!]$).

Let $\Pi = \{\gamma \in \mathcal{L} \mid \{\varphi, \text{♀}, \gamma\} \nvDash \bot\}$. Take any $\gamma \in \Pi$; there is some model $\mathcal{M}_\gamma = (W_\gamma, R_\gamma, \xi_\gamma)$ with $u_\gamma \in W_\gamma$ and $\mathcal{M}_\gamma, u_\gamma \vDash \bigwedge\{\varphi, \text{♀}, \gamma\}$. In which case, there is also a model $\mathcal{M}$ whose domain is the disjoint union of the domains of the $\mathcal{M}_\gamma$'s plus one extra point u^φ, which sees precisely the u_γ's, and where Bob plays some strategy (doesn't matter which) $s_b \in T_b$. To put it more formally, $\mathcal{M}^\varphi$ is $(W^\varphi, R^\varphi, \xi^\varphi)$, where:

$$\begin{array}{rcl} W^\varphi & = & \{u^\varphi\} \cup \bigcup_{\gamma \in \Pi}(W_\gamma \times \{\gamma\}) \\ (x,\gamma)R^\varphi(y,\delta) & \text{iff} & xR_\gamma y \text{ and } \gamma = \delta \\ u^\varphi R^\varphi(x,\gamma) & \text{iff} & x = u_\gamma \\ \xi^\varphi(u) & = & \begin{cases} \xi_\gamma(u) & \text{if } u \in |\mathcal{M}_\gamma| \\ s_b & \text{otherwise.} \end{cases} \end{array}$$

This simple construction is illustrated in Figure 2.6.

Then certainly $\mathcal{M}^\varphi, u^\varphi \vDash \Box\varphi$, and for all $\gamma \in \Pi$, $\mathcal{M}^\varphi, u^\varphi \vDash \Diamond\gamma$. Let $\Gamma^\varphi = \{\psi \in \mathcal{L} \mid \mathcal{M}^\varphi, u^\varphi \vDash \psi\}$, the modal theory of u^φ. Clearly $\Gamma^\varphi \in \textsc{mss}\ (\mathcal{L})$. We want to show that $R(\Gamma^\varphi) = \{\Delta \in \textsc{mss}\ (\mathcal{L}) \mid \varphi \in \Delta\}$.

$\subseteq$: $\Box\varphi \in \Gamma$, so by the Truth Lemma, $\mathcal{C}, \Gamma \vDash \Box\varphi$. In which case we have $\forall \Delta \in R(\Gamma)\ \ \mathcal{C}, \Delta \vDash \varphi$. But again by the Truth Lemma, this means that $\varphi \in \Delta$.

$\supseteq$: Take any $\Delta \in \textsc{mss}\ (\mathcal{L})$ such that $\varphi \in \Delta$. Then take any $\delta \in \Delta$; clearly we have $\{\varphi, \delta\} \nvDash \bot$, so $\delta \in \Pi$, in which case $\mathcal{M}^\varphi, u^\varphi \vDash \Diamond\delta$, i.e. $\Diamond\delta \in \Gamma$. Since δ was arbitrary, then by the definition of R (part of the model $\mathcal{C}$), we have shown that $\Gamma R \Delta$.

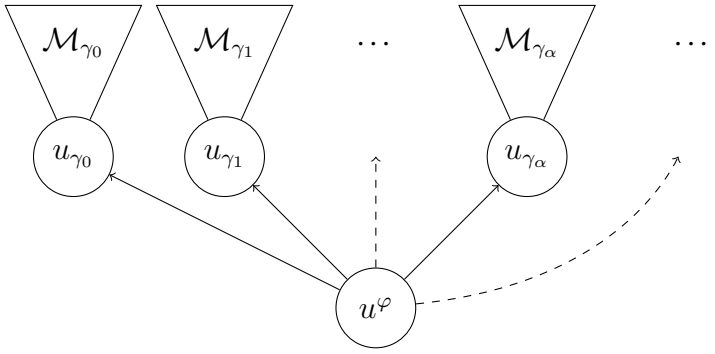


Figure 2.6: The construction of $\mathcal{M}^\varphi$, from the proof of Theorem 2.4

(We initially proved in [Zvesper and Pacuit, 2010] a result for a finitary modal language by giving a completeness proof for the relevant logic, which simply involved adding axioms for ♀ to the known logic for the normal modality $\Diamond$. It might be possible to adapt completeness proofs for infinitary modal logics, but this is not necessary.) ■

This possibility Theorem 2.4 says that *some* of the analyst's tools can be available to the players. Specifically, it shows that the modal language is not too powerful, and thus *can* be 'available' to the players being modelled. We take this to be an avatar of the slogan quoted above from [Blackburn *et al.*, 2001] about locality.

There is a close connection between the failure of assumption-completeness and the impossibility of an unrestricted Comprehension axiom schema in set theory (cf. [Devlin, 1993]). Russell's paradox shows that the formula $x \notin x$ cannot occur in a Comprehension axiom, on pain of inconsistency: it cannot be used to define a set. This connection between Russell's paradox and the "paradox" of assumption-completeness is already remarked in [Brandenburger and Keisler, 2006]. To make that connection precise it suffices to make the following remark: The consistency of a comprehension schema $\exists x \forall z(z \in x \equiv \varphi(z))$, for all $\varphi \in \mathcal{L}$ (with x not free in φ), is essentially a single-agent assumption-completeness for the language defined by $\mathcal{L}$.

[Baltag, 1998] proves a result in the context of non-wellfounded set theory that is therefore related to Theorem 2.4, namely that in that context, an axiom of comprehension for infinitary modal formulae holds. There it is also shown, given some large cardinal assumption, that an axiom of comprehension for so-called 'generalised positive formulae' is consistent. Generalised positive formulae are those defined by the following recursive scheme and having one free variable:

$$\varphi ::= ♀(x) \mid \mathbf{s}_i(x) \mid x\mathbf{R}y \mid \bigwedge \Phi \mid \bigvee \Phi \mid \forall x\varphi \mid \exists x\varphi \mid \forall x(yRx \to \varphi) \mid \forall x(\psi \to \varphi),$$

where ψ *can also include negation*. It would be of interest to pursue this connection with results about the comprehension scheme further, especially in view of [Brandenburger and Keisler, 2006, Theorem 10.4], stating the assumption-completeness of a

certain *positive* (negation-free) fragment of the first-order language, that is related for example to known results in set theory, cf. [Forti and Hinnion, 1989].

What about further strengthening our possibility result? We will look briefly at perspectives for doing exactly this.

In the details of the proof of Theorem 2.4, the only 'modal' behaviour we exploit, in showing the satisfiability of a certain set of sentences, is that the truth of modal sentences is preserved under disjoint unions and generated sub-models. So although we do not have a proof of an analogue for Theorem 2.4 for a language with the binder $\downarrow$, (because the same MSS -based construction as that used in the proof of Theorem 2.4 would not work for $\mathcal{L}_{N,\downarrow}$, in the sense that the Truth Lemma would fail), still we are lead to suspect that this language also has assumption-complete models:

Conjecture 2.1. *For* any *game G, there are full assumption-complete models for $\mathcal{L}^{\Psi}_{\Diamond,O,\downarrow}$ where for every strategy s_i there is a proposition letter $\mathbf{s_i} \in \Psi$.*

An inspection of the proof of Theorem 2.3 reveals three conditions that *together are sufficient* to show that a language $\mathcal{L}$ is not assumption-complete.

The first condition is that $\mathcal{L}$ can express that Ann believes that Bob's assumption is wrong.

$$D_{\mathcal{M}} := \{u \in W_a \mid \forall v \in R_a(u),\ u \notin R_b(v)\}$$

(C1) $\exists \varphi \in \mathcal{L} :\ \forall \mathcal{M},\ [\![\varphi]\!]_{\mathcal{M}} = D_{\mathcal{M}}$.

The second condition is that $\mathcal{L}$ be closed under the assumption operator, i.e.:

(C2) $\varphi \in \mathcal{L} \Rightarrow \exists \psi \in \mathcal{L} : \forall \mathcal{M},\ [\![\psi]\!]_{\mathcal{M}} = \{u \in W_a \mid R_a(u) = [\![\varphi]\!]_{\mathcal{M}} \cap W_b\}$

If both C1 and C2 hold, then $\mathcal{L}$ is assumption-*in*complete. In particular, $\mathcal{L}$ will allow sentence (BK) to be expressed. Note that (C1) holds with respect to $\mathcal{L}_{\Diamond,\downarrow}$:

Fact 2.15. $D_{\mathcal{M}} = [\![\varphi \wedge \downarrow x.\ \Box\, \Box \neg x]\!]_{\mathcal{M}}$

However, since the language is closed under generated sub-models then, importantly, the assumption operator is *not* expressible:

Fact 2.16. *$\mathcal{L}_{\downarrow}$ does not satisfy condition C2.*

Evidence *against* our Conjecture 2.1 is available from the observation that the comprehension scheme in set theory is inconsistent with some $\mathcal{L}_{\downarrow}$ sentences. Indeed, Russell's paradox still has its full force, since $x \notin x$ is in that language ($\downarrow x.\ \Box\, \neg x$). However, the absence of any way to talk about *precisely* one set of the *other* type does lead us still to support Conjecture 2.1. We find it interesting, and worth future research, that there might be such a language for which comprehension in set theory fails, but which is assumption-complete. That would in our view truly highlight the *interactive* nature of assumption-completeness.

We've looked at a language with C1; so what about C2? The easiest way to get C2 is just to add an assumption operator, $\bowtie \varphi$, meaning just that the player's assumption is

φ (cf. [Levesque, 1990; Halpern and Lakemeyer, 2001]). Clearly, adding an assumption operator to $\mathcal{L}_{\downarrow}$ will make the language assumption-incomplete, since it will then have (C1) and (C2). (So this would be a strictly weaker language than first-order logic, that is nonetheless assumption-incomplete.) Nonetheless, while we do not investigate the matter further here, we conjecture that adding an assumption operator into $\mathcal{ML}$ would *not* leave the happy realm of assumption-completeness.

Summary

More issues have been left unresolved than resolved by this Chapter, which we take to be an indication that while many game-theorists have become interested in logic, and logicians in game theory, there is still progress to be made. Concerning the definability of rationality, we made some first steps by showing some languages to be expressive enough. We know of very little work in this direction.

There is also scope, in our view, for philosophical arguments to be tightened in the game-theoretical literature. For example, while we find the concept of assumption-completeness technically elegant, we do not find convincing the arguments that it is somehow intuitively justified because it corresponds to some natural notion of availability of the language to the players. Brandenburger and Keisler also cited a technical 'need' for it, or similar notions, to give conditions for backward induction [Battigalli and Siniscalchi, 2002] or iterated admissibility [Brandenburger *et al.*, 2008]. Yet we will present later our own condition for backward induction, in Chapter 4, that does not require any form of 'complete' model. For the case of admissibility we do have little to say (though we will consider some epistemic aspects of it in the next Chapter). Still, it is questionable how informative it is to say that the epistemic conditions for playing according to the iterated elimination of weakly dominated strategies are somehow elucidated by saying that the players are 'in' a complete belief model. Nonetheless, we gave one substantial technical contribution in this Chapter, which was to prove that infinitary modal languages *are* assumption-complete.

This was a language, recall from our extensive cataloguing, that can express common belief and rationality, two concepts that are central to game theory. In that catalogue of languages, we discussed various ways in which to define some notions, like optimality and rationality, that are relevant to the analysis of games.

Subscript	Symbols and their interpretations
N	$(W, \mathcal{N}_i, \ldots)_{i\in N}, u \vDash \Box_j\varphi$ iff $[\![\varphi]\!] \in \mathcal{N}_j(u)$ $(W, R_i, \ldots)_{i\in N}, u \vDash \Box_j\varphi$ iff $R_j(u) \subseteq [\![\varphi]\!]$ Belief modalities
$\bigcirc$	$(W, O_i, \ldots)_{i\in N}, u \vDash \bigcirc_j\varphi$ iff $u \in O_j([\![\varphi]\!])$ 'Optimality' modalities
κ	$\mathcal{M}, u \vDash \bigwedge_{\beta\in\alpha} \varphi_\beta$ iff $\forall\beta \in \alpha$, $\mathcal{M}, u \vDash \varphi_\beta$ Conjunctions of length $< \kappa$
i_s	$\mathcal{M}, u \vDash [i_s]\varphi$ iff for all v, if $\xi_i(u) = \xi_i(v)$ then $\mathcal{M}, u \vDash \varphi$ Strategy modalities
A	$\mathcal{M}, u \vDash A\varphi$ iff $[\![\varphi]\!] = \lvert\mathcal{M}\rvert$ Global modality
$\geq, >, \leq, <$	$\mathcal{M}, u \vDash \langle\geq_i\rangle\varphi$ iff $\exists s_i \in \xi_i([\![\varphi]\!]_\mathcal{M}) : \xi_i(u) \geq_i s_i$ Preference modalities
$\forall$	$\mathcal{M}, u \vDash_\delta \forall x.\varphi$ iff for all $\delta' \sim_{-x} \delta$, $\mathcal{M}, u \vDash_{\delta'} \varphi$ First-order quantifiers
@	$\mathcal{M}, u \vDash_\delta @_x\varphi$ iff $\delta(x) = \{v\}$ & $\mathcal{M}, v \vDash_\delta \varphi$ Hybrid modalities
$\downarrow$	$\mathcal{M}, u \vDash_\delta \downarrow x.\varphi$ iff $\mathcal{M}, u \vDash_{\delta[x\mapsto\{u\}]} \varphi$, Bounded first-order quantification
$\overline{\forall}$	$\mathcal{M}, u \vDash_\delta X$ iff $u \in \delta(X)$ $\mathcal{M}, u \vDash_\delta \overline{\forall}X\varphi$ iff for all $\delta' \sim_{-X} \delta$, $\mathcal{M}, u \vDash_{\delta'}$ Propositional ('second-order') quantifiers
ν	$\mathcal{M}, u \vDash_\delta \nu X.\varphi$ iff $\exists E \subseteq \lvert\mathcal{M}\rvert : u \in E \subseteq [\![\varphi]\!]^{\delta[X\mapsto E]}$ Fixpoint quantifiers
-1	$(W, R_i, \ldots)_{i\in N}, u \vDash \Box_j^{-1}\varphi$ iff $R_j^{-1}(u) \subseteq [\![\varphi]\!]$ Inverse belief modalities

For reference, we provide a table listing the various components of languages that we considered. In the right-hand column we give the formal semantics of various symbols, and a reminder of what we called the symbols, and in the corresponding left-hand column is the subscript we used to denote that those symbols are in a language.

Chapter 3

Dynamics

"*Public announcements nurture me as I grow towards my time*".
"*Public announcements have punctuated my life*".
– Saleem [Rushdie, 1981]

Let us recap two observations from Chapter 1, that we did not elaborate upon there. Firstly, there is something suspect about the claimed 'equivalence' between a given epistemic condition and its consequence O^∞. And secondly, in case the optimality operator O was not monotonic, there was a problem even with specifying what the condition should be, that entails O^∞. Part of what we do in this Chapter is related to these two concerns: we turn our attention to *where the models come from*, in the sense of looking at formalising some process of deliberation that captures some basic intuitions about the one-shot interaction that forms the basis of our Deductive interpretation of game theory.

Always in the spirit of the previous Chapter and its emphasis on logical syntax, we will therefore present some existing work on what is known as 'dynamic epistemic logic' (DEL). Throughout this Thesis so far, we have used neighbourhood models, whereas mainstream epistemic logic has usually used relational models. That holds for the DEL too, and one minor technical contribution of this Chapter is to show how DEL can be generalised to neighbourhood models.

Partly our concern here with dynamics of logical systems is for its own sake, and partly it is to complete the picture of those aspects that we have studied in Chapter 1 of the deductive interpretation of strategic games.

> "*Logical systems as they stand are product-oriented, but Logical Dynamics says that both sides of the duality should be studied to get the complete picture.*" – [Benthem, forthcoming]

We flesh out an idea given in passing in [Benthem, 2007b], that dynamic epistemic logic can be used to reason about where the epistemic models that we described in Chapter 1 come from.

The idea that we will play with is the following. To begin with, take a 'blank' model of the game, in which in some sense nothing is believed. This model is not a 'complete' belief model of some kind, but a very simple model that faithfully represents the epistemic situation in a one-shot interaction *before* the players have started deliberating. Then add beliefs to the players, in some systematic way, until a certain configuration of beliefs is obtained, that respects some epistemic condition. Finally, another epistemic action corresponds to that of the players each choosing the strategy they will play.

The "in some systematic way" can be fleshed out differently, but revolves around the idea of what is known as a "public announcement". We suggest different interpretations stating how this epistemic action of public announcement should be understood, but the one we will favour is that it represents some kind of "private but common" reasoning process, and so is in tune with the deductive interpretation of game theory. We do not find here an application for many of the subtleties of analysis offered by existing dynamic epistemic logics which, as we will mention, are claimed by some to be in a certain sense "complete" for the epistemics of social or interactive situations. This recommends them for further application in game theory.

An interesting feature of the public announcements involved in the process we described is that the announcement that is repeated is *syntactically* the same: it is an announcement *that the players are rational*. Thus one repeatedly announces *the same* (syntactic) sentence or formula, but the (semantic) event that is its meaning *changes each time it is announced*. This feature is exhibited in other examples for which DEL provides elegant analyses, most notably the 'muddy children' example analysed for example in [Plaza, 1989]. In that example the story is that a group of n children have been playing in a garden, with the result that $k \leq n$ of them have muddy foreheads. And adult arrives and informs them that at least one has a muddy forehead. She then repeatedly asks for a 'show of hands' from children who know whether they are muddy. Since this show of hands is supposed to be simultaneous, we could model it as a single public announcement of the conjunction of all the individual announcements. Take the case where $k = n = 3$. Then the first show of hands says: 'none of the children know whether they are muddy'. *So does the second*. And yet this second announcements conveys, in the context of the previous announcement, new information, so that all children (assuming that it is common true belief amongst them that they are all perfect reasoners) will after these announcements know that they have muddy foreheads.

Separately from all of that, we will also use another kind of model for belief, *plausibility models*, that allow for the representation of *conditional beliefs*. The conditional beliefs of a player specify not just the things that she believes are true, but also those things she would believe if she were to learn that her actual beliefs were incorrect. We will show how conditional beliefs are important in understanding non-monotonic optimality operators, by arguing that relational or neighbourhood models are not adequate for providing an explanation of *why* players would play the strategies indicated by the solution algorithm. That is, in the case of non-monotonic procedures, the models obtained by the epistemic actions we just mentioned are not *stable*, or *self-enforcing*; they

do not contain what we will call a '*rational equilibrium of beliefs*'.

In the context of plausibility models, there is another kind of public announcement, that models so-called 'soft information' (to use the terminology of [Benthem, 2007a]), as opposed to the 'hard information' of the more standard announcement. Iterated 'soft' announcements, of what we call 'lexicographic rationality' from an appropriately blank initial plausibility model, will lead to a model in which there *is* rational equilibrium of beliefs. This goes some way towards explaining why the players would play according to the algorithm of iterated elimination of non-optimal strategies, even in some cases where the optimality operator is non-monotonic.

Background literature

Dynamic epistemic logic started with [Plaza, 1989; Gerbrandy and Groeneveld, 1997; Gerbrandy, 1999; Baltag *et al.*, 1999], and is presented in the textbook [Ditmarsch *et al.*, 2007]. The results we present on neighbourhood semantics are related to [Rodriguez, 2007].

We have mentioned that we take some ideas from [Benthem, 2007b], but we should make it clear that our considerations here are relatively superficial: that paper addresses a number of deep questions involving logical definability that we do not touch upon here.

Conditional beliefs and plausibility models (also sometimes called 'conditional doxastic models') are developed in [Board, 2002; Benthem, 2007a; Baltag and Smets, 2006; 2008b].

On a less related note: an epistemic foundation for the iterated elimination of weakly dominated strategies was provided in [Brandenburger *et al.*, 2008], and while we do not pretend to give an epistemic foundation here, it is worth mentioning that the idea of using conditional belief models to understand the corresponding non-monotonic optimality operator is already present there in the notion of a lexicographic probability system [Blume *et al.*, 1991].

Organisation of the Chapter

Section 3.1 contains essentially no discussion of game theory, but focuses on the logical aspects of the dynamics of belief. In particular, we will present there reduction axioms for a number of different kinds of action. First of all we look at public announcements, and show (Propositions 3.1–3.4) that known results about completeness carry over to neighbourhood semantics. Next we look at the more general epistemic actions from [Baltag *et al.*, 1999], and our contribution there is to show completeness of a modal language with action modalities for a neighbourhood semantics. Finally, we introduce a new announcement operator that adds information but also can introduce 'ignorance about ignorance', and indeed turn a neighbourhood model that is relational into one that is not.

That announcement operator is the one that we use at the end of Section 3.2 in order to show where the model in Theorem 1.5 comes from. However, it is not very intuitive, and the earlier parts of the Section examine more intuitive announcement operators on strategic game models. In Section 3.3 we give a brief exposition of plausibility models and conditional beliefs. That will be used later in Chapter 4, but we introduce it in this Chapter because in the same Section we will introduce the notion of rational equilibrium of beliefs, and argue for the use of plausibility models in epistemic analysis of non-monotonic solution concepts.

3.1 Dynamic epistemic logic

Dynamic epistemic logic (DEL) is the logic of models of belief as they change under the effect of various kinds of information flow. In this Section we will go through existing results about dynamic epistemic logic, starting with the simple logic of public announcements, and then looking at 'update models'. Thus rather than look just at a single model, we will be interested in how models change, and what happens when they do change, as the players being modelled acquire information. In the first two parts of the Section the contribution that we make is to generalise existing results from relational models, on which DEL has usually been studied, to neighbourhood models.

In the last part of this Section we introduce a new kind of public announcement operator. This operator creates beliefs while at the same time creating 'ignorance about ignorance', to borrow an expression from [Samet, 1990]. As we will show in Section 3.2, this new operator is the one that is needed in order to generate the model described in Theorem 1.5.

Public announcements

A *public announcement* is perhaps the simplest kind of information change in a group. In its most natural formalisation, a public announcement of A eliminates from the model all those states in which A is false. The particular kind of information acquisition that we consider here is the 'public announcement' kind, that is one way to represent all players synchronously learning some piece of information. The 'public announcement' metaphor need not be taken literally: it can also be a kind of joint discovery of any kind. We discuss various interpretations, and look at variants of public announcements, later in this Chapter. What is essential to public announcements is that they are a *collective* action of 'learning'.

Whether we should call this an 'action' is questionable, since that noun naturally implicates some form of agency, that is absent from the formalism here. However, the obvious alternative, that is taken in the philosophical literature [Davidson, 1980] to indicate an agency-less 'action', is 'event', which is sadly already taken. We prefer to avoid the ambiguity, and so use the word 'action' for what is sometimes in the dynamic epistemic literature logic called 'event'.

We have in general shied away from calling anything 'knowledge', but in the case of the kind of public announcement that we will consider in this Section, we will generally consider that they *do* generate knowledge. This kind of announcement models what van Benthem [Benthem, 2007a] calls 'hard information', and can be thought of as an act of observation of some absolute fact. If our models are to be able to give an interpretation to knowledge, then surely we would want at least that the tautological event (the whole state space) should be something the players 'know' themselves to be in. Thus if announcements reduce the state space, they would generate knowledge. We will not explicitly introduce knowledge into the language until we look in Section 3.3 at conditional belief models, where we will be able to introduce a technically and conceptually significant difference between belief and knowledge. For the time being, just note that it would be possible to introduce a knowledge operator, and that everything we say regarding the belief operator and public announcements would also apply to a knowledge operator.

On the level of the language, we can add a binary modality $\langle !\varphi\rangle\psi$. So given a language $\mathcal{L}$, we turn it into a language $\mathcal{L}_!$ with hard public announcements by adding the following clause:

$$\varphi ::= \ldots \mid \langle !\varphi\rangle\varphi.$$

The semantics of the public announcement operator are given in terms of '*relativisation*'. The intuitive idea is that a public announcement of φ rules out entirely and irrevocably, for everybody, all states in the model $\mathcal{M}$ at which φ does not hold. Since the action is 'public', or 'collective', the information change brought about by it is faithfully represented by eliminating all states at which φ does not hold, or to put it another way: by *relativising* to $[\![\varphi]\!]_{\mathcal{M}}$.

Definition 3.1. Given a model $\mathcal{M} = (W, \mathcal{N}_i, \xi)_{i\in N}$, the ***relativisation*** to $A \subseteq W$ is just the model $\mathcal{M}!A = (A, \mathcal{N}_i!A, \xi!A)_{i\in N}$ where $\mathcal{N}_i!A$ and $\xi!A$ are as follows, with domain A:

$$\begin{array}{ll} \mathcal{N}_i!A(u) & = \{U \cap A \mid U \in \mathcal{N}_i(u)\} \\ \xi!A(u) & = \xi(u). \end{array}$$

It is convenient to write $\mathcal{M}!\varphi$ for $\mathcal{M}![\![\varphi]\!]_{\mathcal{M}}$. Then the semantic clause for the public announcement modality is given in terms of relativisation:

$$\mathcal{M}, u \Vdash \langle !\varphi\rangle\psi \quad \mathcal{M}, u \Vdash \varphi \;\&\; \mathcal{M}!\varphi, u \Vdash \psi.$$

Public announcement modalities are already definable in many of the languages that we have considered, since model-theoretically speaking they are closed for relativisation. That is to say, for many of the languages $\mathcal{L}$, given a formula φ in $\mathcal{L}_!$, there is a formula ψ in $\mathcal{L}$ such that on any model $\mathcal{M}$, $[\![\varphi]\!]_{\mathcal{M}} = [\![\psi]\!]_{\mathcal{M}}$. In order to show this property, we can give a recursive translation from $\mathcal{L}_!$ to $\mathcal{L}$. For example, a translation like the following one was in effect given in [Plaza, 1989], for the language $\mathcal{L}^{\Psi}_{N,O,\kappa,!}$. (Plaza considered the finitary case where $\kappa = \aleph_0$ and considered relational, rather than

arbitrary monotonic, modal operators; the extension to cover the more general cases is straightforward.)

$$\begin{array}{lcl}
\mathrm{tr}(\langle !A\rangle \mathbf{s_i}) & = & A \wedge \mathbf{s_i} \text{ for } s_i \in \Psi \\
\mathrm{tr}(\langle !A\rangle \neg\varphi) & = & A \wedge \neg \mathrm{tr}(\langle !A\rangle \varphi) \\
\mathrm{tr}(\langle !A\rangle \bigwedge \Phi) & = & \bigwedge\{\mathrm{tr}(\langle !A\rangle \varphi) \mid \varphi \in \Phi\} \\
\mathrm{tr}(\langle !A\rangle \Box_i \varphi) & = & A \wedge (\Box_i(A \rightarrow \mathrm{tr}(\langle !A\rangle \varphi))) \\
\mathrm{tr}(\langle !A\rangle \bigcirc_i \varphi) & = & A \wedge (\bigcirc_i(A \rightarrow \mathrm{tr}(\langle !A\rangle \varphi))) \\
\mathrm{tr}(p) & = & p \\
\mathrm{tr}(\bigwedge \Phi) & = & \bigwedge\{\mathrm{tr}(\varphi) \mid \varphi \in \Phi\} \\
\mathrm{tr}(\neg\varphi) & = & \neg \mathrm{tr}(\varphi) \\
\mathrm{tr}(\Diamond\varphi) & = & \Diamond \mathrm{tr}(\varphi).
\end{array}$$

This translation removes all occurrences of $\langle !A\rangle$ from any given formula, thus it is indeed a translation from $\mathcal{L}_{N,\kappa,!}$ to $\mathcal{L}_{N,\kappa}$. Therefore it suffices, in order to prove that there is a formula in $\mathcal{L}$ equivalent to any in $\mathcal{L}_!$, to show that $[\![\mathrm{tr}(\varphi)]\!] = [\![\varphi]\!]$. And indeed, *on monotonic models*, the translation $\mathrm{tr}(\cdot)$ we just gave does preserve truth:

Proposition 3.1. $\mathrm{tr}(\cdot)$ *preserves truth on monotonic models. That is: for any monotonic model* $\mathcal{M}$*, and any* $\varphi \in \mathcal{L}_{N,\kappa,!}$*,* $[\![\mathrm{tr}(\varphi)]\!] = [\![\varphi]\!]$.

Proof. By induction on φ, for a suitably defined notion of the complexity of formulae of the language. The only interesting step is that for $\varphi := \langle !A\rangle \Box_i \psi$. We give this here because while the relational version is known, the more general monotonic neighbourhood result is new. So suppose (a consequence of the inductive hypothesis) that on all models $\mathcal{M}$:

$$[\![\mathrm{tr}(\langle !A\rangle \psi)]\!]_{\mathcal{M}} = [\![\langle !A\rangle \psi]\!]_{\mathcal{M}} = [\![A]\!]_{\mathcal{M}} \cap [\![\psi]\!]_{\mathcal{M}!A}.$$

Take any model $\mathcal{M}$. We want to show that $[\![\mathrm{tr}(\langle !A\rangle \Box_i \psi)]\!]_{\mathcal{M}} = [\![\langle !A\rangle \Box_i \psi]\!]_{\mathcal{M}}$. That is established by the following equivalences:

$$\begin{array}{lrcl}
\mathcal{M}, u \Vdash \langle !A\rangle \Box_i \psi & \Leftrightarrow & \mathcal{M}, u \Vdash A \text{ and } & \mathcal{M}!A, u \Vdash \Box_i \psi \\
& & \Leftrightarrow & [\![\psi]\!]_{\mathcal{M}!A} \in \mathcal{N}!A_i(u) \\
& & \Leftrightarrow & [\![\langle !A\rangle \psi]\!]_{\mathcal{M}} \in \mathcal{N}!A_i(u) \\
& & (I.H.) \Leftrightarrow & \exists X \in \mathcal{N}_i(u) : X \cap [\![A]\!]_{\mathcal{M}} = [\![\langle !A\rangle \psi]\!]_{\mathcal{M}} \\
& & (Mon) \Leftrightarrow & [\![\langle !A\rangle \psi]\!]_{\mathcal{M}} \cup [\![\neg A]\!]_{\mathcal{M}} \in \mathcal{N}_i(u) \\
& & \Leftrightarrow & \mathcal{M}, u \Vdash \Box_i(\neg A \vee \langle !A\rangle \psi) \\
& \Leftrightarrow & \mathcal{M}, u \Vdash A \ \wedge & \Box_i(A \rightarrow \langle !A\rangle \psi).
\end{array}$$

■

It is easy to see that the monotonicity condition is necessary:

Fact 3.1. *There is a (non-monotonic) neighbourhood model* $\mathcal{M}$ *and a formula in* $\mathcal{L}_{!,N}$*such that,* $[\![\mathrm{tr}(\varphi)]\!]_{\mathcal{M}} \neq [\![\varphi]\!]_{\mathcal{M}}$.

Proof. Let $\mathcal{M} = (\{a, b\}, \mathcal{N}, \xi)$, $\mathcal{N}(b) = \{b\}$ and $\xi(a) \neq \xi(b) = s$. Set $\varphi = \langle !s \rangle \Box_i s$. Then $\mathrm{tr}(\varphi) = s \wedge \Box_i(s \to s)$. Notice then that $\mathcal{M}, a \nVdash \mathrm{tr}(\varphi)$, since $[\![s \to s]\!]_{\mathcal{M}} = \{a, b\} \notin \mathcal{N}(a)$. However, $\mathcal{M}, b \Vdash \mathrm{s}$ and $\mathcal{M}!p, b \Vdash \Box_i \mathrm{s}$, so by the definition of the semantics of $\langle !p \rangle$, $\mathcal{M}, b \Vdash \varphi$. ■

The correctness of the translation above is equivalent to the following formulae being valid:

$$\begin{array}{lcl} \langle !A \rangle p & \equiv & A \wedge p \\ \langle !A \rangle \neg\varphi & \equiv & A \wedge \neg\langle !A \rangle \varphi \\ \langle !A \rangle \bigwedge \Phi & \equiv & \bigwedge \{ \langle !A \rangle \varphi \mid \varphi \in \Phi \} \\ \langle !A \rangle \Box_i \varphi & \equiv & A \wedge (\Box_i (A \to \langle !A \rangle \varphi)) \\ \langle !A \rangle \bigcirc_i \varphi & \equiv & A \wedge (\bigcirc_i (A \to \langle !A \rangle \varphi)) \end{array}$$

Axiomatically speaking, these validities can be used as so-called ***reduction axioms*** to prove completeness of the language $\mathcal{L}_{N,!}$ by reducing it to the language $\mathcal{L}_N$. In giving other forms of epistemic action in what follows, for notational ease we prefer to list the reduction axioms than the resulting translation.

What about the case of non-monotonic modalities? The public announcement logic of non-monotonic modalities, as we will now see, can be treated in the same was as the public announcement logic of common belief. [Benthem *et al.*, 2005] consider the question of adding epistemic action modalities to a logic for common knowledge, and show that while it is *not* possible to reduce a language with a common knowledge modality and public announcement modalities to the language without public announcement modalities, there *is* such a reduction with respect to a language including also a *relativised common knowledge* modality. Thus they define, for relational models, a binary modal operator

$$\varphi ::= \ldots \mid \Box^{\infty}(\varphi, \psi),$$

Proposition 3.2. *In the context of neighbourhood semantics, the semantic clause for* $\Box^{\infty}(\varphi, \psi)$ *given in [Benthem* et al.*, 2005] is equivalent to the following, where* $\mathcal{M} = (W, \mathcal{N}_i, \xi)_{i \in N}$*:*

$$\mathcal{M}, u \Vdash \Box^{\infty}(\varphi, \psi) \quad \text{iff} \quad \exists E \subseteq W : u \in E \;\&\; \forall x \in E \quad \begin{array}{l} E \in \mathcal{N}![\![\varphi]\!]_{\mathcal{M}}(x) \\ \text{and} \quad [\![\psi]\!]_{\mathcal{M}} \in \mathcal{N}![\![\varphi]\!]_{\mathcal{M}}(x). \end{array}$$

[Benthem *et al.*, 2005] give a valid reduction axiom for this new binary operator, that can be trivially extended to the case of belief:

Proposition 3.3. *The following reduction axiom is valid on neighbourhood models:*

$$\langle !A \rangle \Box^{\infty}(\varphi, \psi) \equiv A \wedge \Box^{\infty}(\langle !A \rangle \varphi, [!A]\psi).$$

This modality $\Box^{\infty}(\varphi, \psi)$ is not expressible in the basic modal language $\mathcal{L}_N$ [Benthem *et al.*, 2005, Theorem 1]. It manages to make the language closed for relativisation by 'pre-encoding' of all the possible relativisations into the original language. So

the same trick can be used in order to cover the case of non-monotonic modalities.[1] Given some arbitrary neighbourhood modality $\bigcirc\varphi$, whose neighbourhood function $\mathcal{N}$ might not be monotonic, we can similarly specify the binary, relativised, version $\bigcirc(\varphi, \psi)$ with the following semantics:

$$\mathcal{M}, u \Vdash \bigcirc(\varphi, \psi) \quad \text{iff} \quad [\![\psi]\!]_{\mathcal{M}} \in \mathcal{N}![\![\varphi]\!]_{\mathcal{M}}(u).$$

As we will see in the next Section, we might also want to consider a simple modification of this kind of public announcement, that is studied in [Benthem and Liu, 2007]. Public announcements add beliefs. Usually this is done by eliminating states from the model. (This means among other things that the public announcement has also to be *true*.) A slight variant of this kind of state-eliminating public announcement changes instead the belief component of the model, actually in precisely the same way as in the state-eliminating version, but does not eliminate the relevant states. So in the case of relational models, this means cutting links, not removing worlds. In the general monotonic neighbourhood semantics it means intersecting neighbourhoods, but not removing the relevant states from the model. We call this *non-eliminative* (hard) announcement, and define it as follows. $(W, \mathcal{N}_i, V)_{i\in N}¡A = (W, \mathcal{N}'_i, V)_{i\in N}$ with

$$\mathcal{N}'_i(u) = \{A \cap E \mid E \in \mathcal{N}_i(u)\}.$$

Non-eliminative announcement is also easily reduced with respect to the basic modal language; the relevant reduction axioms are given in Proposition 3.4, which entails that there is a straightforward translation from $\mathcal{L}_{N,O,\kappa,¡}$ to $\mathcal{L}_{N,O,\kappa}$.

Proposition 3.4. *The following reduction axioms are valid on monotonic models.*

$$\begin{array}{rcl}
\langle ¡A\rangle p & \equiv & p \\
\langle ¡A\rangle \neg\varphi & \equiv & \neg\langle ¡A\rangle\varphi \\
\langle ¡A\rangle \bigwedge \Phi & \equiv & \bigwedge\{\langle ¡A\rangle\varphi \mid \varphi \in \Phi\} \\
\langle ¡A\rangle \Diamond_i\varphi & \equiv & \Diamond_i(A \wedge \langle ¡A\rangle\varphi) \\
\langle ¡A\rangle \bigcirc_i \varphi & \equiv & \bigcirc_i(¡A\varphi)
\end{array}$$

(Cf. [Benthem and Liu, 2007, Theorem 4.3].) The same issues concerning 'relativised belief', for an axiomatisation of common belief or with respect to non-monotonic models, arise here as in the eliminative version, and find the same resolution as there.

Epistemic actions

In the dynamic epistemic logic studied in [Gerbrandy and Groeneveld, 1997; Baltag *et al.*, 1999], epistemic *actions* take on the same status as epistemic models. Baltag and Moss [2004] propose two Theses, that are spelled out in too much detail to quote them

[1] This observation is due to Johan van Benthem.

here, but effectively claim that relational models can completely describe the epistemic features of any "social situation" (op.cit. p.166), and that similarly (relational) *action models*, that we will define now, also completely describe the relevant features of any "*social 'action'*" (op.cit. p.167).

Actually, since the form of DEL considered in [Baltag and Moss, 2004] does not handle belief *revision*, Baltag and Moss' Theses themselves as they stand need revising. *Revising* one's beliefs, upon acquiring 'surprising' information that contradicts one's theory of how the world is, is surely an important part of social interaction, but is excluded by the DEL considered in that paper. So the thesis as it stands could maybe be applied to the richer formalism of conditional doxastic models developed in [Baltag and Smets, 2008a]. We will discuss logics for belief revision below in Section 3.3; for now let us assume some variation of the thesis, that excludes revisable belief, which cannot be dealt with by the DEL we present in this Section. Still, we'll refer to the claim that DEL is sufficient to represent all social situations and actions as the "BM thesis".

DEL generalises the logic of hard public announcements, by allowing many more kinds of epistemic actions, that can be much more complicated than announcements or observations that are commonly observed by all players. In particular, it can deal with actions that involve subgroups of the players receiving information; the players receiving different information; suspecting that other players are receiving information that they are not receiving, and so on.

We devote the rest of this Section to presenting the dynamic epistemic logic framework. As far as we are aware, in existing work it has always been presented in terms of relational semantics. Since we have been interested in the more general *neighbourhood semantics*, we present dynamic epistemic logic in terms of neighbourhoods rather than relations. As with public announcements, we will restrict our attention to *monotonic neighbourhood models* and will develop analogues for all of the existing notions in DEL. We also give reduction axioms for neighbourhood DEL, and prove their correctness. First, let us recall the existing, relational, dynamic epistemic logic.

In order to model all of the different kinds of informational actions that the BM thesis claims are modelled by DEL, it uses so-called "action models", that are supposed to represent actions of an epistemic character, in the same way as relational models (which to disambiguate we will sometimes call "state models") are supposed to represent an epistemic situation. Thus an action model $\mathcal{A}$ will be *applied* to a state model $\mathcal{M}$ via an 'update' operation in order to yield a new state model $\mathcal{M} \otimes \mathcal{A}$, that represents the epistemic situation after the action represented by $\mathcal{A}$ has occurred in the situation represented by $\mathcal{M}$.

We will first of all present the relational version of DEL action models and update. Assume fixed some set N of players. A ***relational action model*** $\mathcal{A} = (\Sigma, \rightarrow\!\!\triangleright_i, \mathrm{PRE})_{i \in N}$ is very much like a relational (state) model, i.e. Σ is a set, that we now call 'actions', and each $\rightarrow\!\!\triangleright_i \subseteq \Sigma \times \Sigma$ is a relation over the actions. This time the interpretation of $d \rightarrow\!\!\triangleright_i e$ is that *if d is occurring, then i thinks that e might be occurring*. Notice that this is very similar to the interpretation of the relation R_i in a state model: uR_iv means

that if the state is u then i considers that the state might be v. The function PRE now associates each action $d \in \Sigma$ with a *pre-condition*, i.e. a formula in some language $\mathcal{L}$: $\text{PRE} : \Sigma \to \mathcal{L}$. So the difference between state models and action models is that where each state in a state model is associated with an outcome of a game, each action in an action model is associated with a single sentence, that could express an outcome, or an epistemic proposition, or indeed anything the language can define.

The idea of the precondition is that it is a condition that $\text{PRE}(e)$ is a necessary condition for the action e to occur. So in some sense $\text{PRE}(e)$ gives the '*meaning*' of e, in that it gives the truth-conditions for it. Before defining the product update operation $\otimes$ that is used to apply an action model to a state model, let us use these informal descriptions of what action models are in order to show that we can capture, intuitively, the action of a public announcement. Public announcements are the simplest possible kind of action model. The action model contains only one action, since all players are commonly aware of what action is taking place: the action model $(\{a\}, \{(a,a)\}, \{(a,\varphi)\})_{i\in N}$, in which there is one action a, with precondition φ, and where all players think a is occurring when it occurs, is what we mean by a public announcement. We illustrate this model in Figure 3.1. Action models can express simple variations of the public

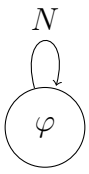


Figure 3.1: An action model for public announcement of φ

announcement operator, for example in the model depicted in Figure 3.2, the action a is an announcement *to the subgroup* $M \subseteq N$ that φ. Notice then that when e happens,

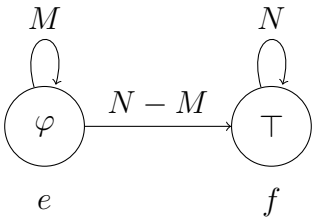


Figure 3.2: A 'private' announcement to the subgroup M

the players in M only consider it possible that e happens, whereas the players *not* in the group M believe that f is happening. And action f is the 'null' action: the action without precondition and where if it happens everybody only considers it possible that it is happening. – It is the 'action that nothing happens'. Therefore the players in M

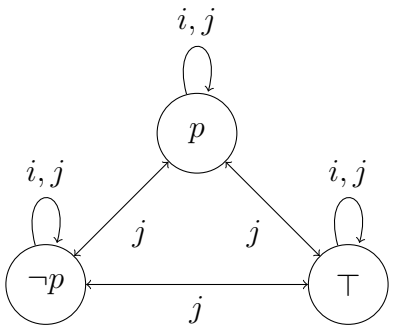


Figure 3.3: An action model representing the 'envelope opening' situation

think, whether e or f actually occur, that f occurs. So when e happens, the players in M learn that φ was true, and also that the players not in M have not learnt anything at all, since they were not *aware* of the announcement taking place.

The action model in Figure 3.3 represents another situation. Here, we can tell a story that fits that model: player i is given an envelope that might contain the result of her recent mathematics examination, or might be some junk mail. She is with a friend, j, when she opens the envelope and looks at the contents, but j does not see what is inside the envelope. Either she has passed (expressed in the language as p) or failed ($\neg p$), and she might find out the result, or might not.

Clearly, we could go on with more examples to back up the BM thesis, but let us return to formal definitions. Given a state model $\mathcal{M} = (W, R_i, \xi)_{i\in N}$ and an action model $\mathcal{A} = (\Sigma, \rightarrowtriangle_i, \text{PRE})_{i\in N}$, the update operation $\otimes$ should be defined in such a way as to capture the intuitive meaning of the relations R_i and $\rightarrowtriangle_i$, and the idea that the action model 'happens' to the state model. I.e. $\mathcal{M} \otimes \mathcal{A}$ should represent the situation after $\mathcal{A}$ has occurred, so that if u was the actual state in $\mathcal{M}$, and e the action in $\mathcal{A}$, then there will be a new state (that will be the pair (u, a)) that is the new actual state.

The ***relational product update*** operation is defined as follows:

$$(W, R_i, \xi)_{i\in N} \otimes (\Sigma, \rightarrowtriangle_i, \text{PRE})_{i\in N} = (W \otimes \Sigma, R_i \otimes \rightarrowtriangle_i, \xi \otimes \Sigma),$$

where

$$\begin{array}{rcl} W \otimes \Sigma & = & \{(u, e) \in W \times \Sigma \mid (W, R_i, \xi)_{i\in N}, u \Vdash \text{PRE}(e)\} \\ (u, d) R_i \otimes \rightarrowtriangle_i (v, e) & \text{iff} & u R_i v \text{ and } d \rightarrowtriangle_i e \\ \xi \otimes \Sigma(u, d) & = & \xi(u). \end{array}$$

Let us take a moment to look at this definition to see that it make sense of our intuitive description. The new states (possibilities) are indeed all the *possible* combinations of previous state and action. And at the state (u, d), i.e. where the state was u, and the action d just occurred, the player i considers (v, e) possible just if she previously considered v possible, and when d occurs, she considers it plausible that e is occurring.

That makes sense for the following reasons: If i realised already that v was not the actual state at u then she will now remember that v was not the actual state, so (v, e) cannot possibly be the actual state. This assumes a notion of 'perfect recall'. Similarly, if she realises (or indeed falsely believes) that e has *not* just happened, then she will not consider it plausible that (v, e) be the new state. Conversely, if she previously considered it possible that v, and considered it possible that e just occurred, then, as long as e is possible at v, she will now consider it possible that (v, e) is the actual state. That is, she does not learn anything 'miraculously', without there being a reason that is given in the action model. Indeed, these notions of 'perfect recall' and 'no miraculous learning' can be formalised, and are shown in [Benthem and Pacuit, 2006] to characterise product update within a broader temporal logic framework.

Just as we added modalities for public announcements to our modal language and showed that they can be reduced, so we can also add modalities $\langle \mathcal{A}, e \rangle \varphi$, meaning 'after e in action model $\mathcal{A}$ occurs, φ holds.' So the language of dynamic epistemic actions $\mathcal{L}_{\otimes}$ is formed from $\mathcal{L}$ by adding the following clause:

$$\varphi ::= \ldots \mid \langle \mathcal{A}, e \rangle \varphi,$$

where $\mathcal{A}$ is an action model, and e is an action in it. In the relational case, the semantics is given in terms of relational update:

$$\mathcal{M}, u \Vdash \langle \mathcal{A}, e \rangle \varphi \quad \text{iff} \quad \mathcal{M} \otimes \mathcal{A}, (u, e) \Vdash \varphi.$$

And, again in the public announcement vein, as long as $\kappa > \#(\mathcal{A})$, 'reduction axioms' can be shown to be valid, thereby allowing a translation from the language $\mathcal{L}_{N,\kappa,\mathcal{A}}$ to $\mathcal{L}_{N,\kappa}$.

Proposition 3.5 ([Baltag *et al.*, 1999, Proposition 3.1])**.** *The following reduction axiom is valid for relational update:*

$$\langle \mathcal{A}, e \rangle \, \Box_i \, \varphi \equiv \text{PRE}(e) \wedge \bigwedge_{d \leftarrow_i e} \Box_i([d]\varphi)$$

Note that we are dealing here with the *eliminative* version, in which not only are links cut to states that are ruled out, but also those states are removed entirely from the model. It is also possible to specify a non-eliminative version of product update, along similar lines to non-eliminative announcements, but we leave the details for the reader to fill in.

Let us turn our attention now to generalising relational product update to product update for (monotonic) neighbourhood models. We again will only consider monotonic models, since it will not be possible to give reduction axioms for the other case.

So given a monotonic neighbourhood model $\mathcal{M} = (W, \mathcal{N}_i, \xi)_{i \in N}$, and a neighbourhood action model $\mathcal{A} = (\Sigma, \mathcal{E}_i, \text{PRE})_{i \in N}$, we want to define the ***neighbourhood product update*** $\mathcal{M} \boxtimes \mathcal{A}$. We again consider the eliminative version, so the new state

space is still $W \otimes \Sigma$ as above, and the new outcome function remains $\xi \otimes \Sigma$. The difference of course is in the definition of the new neighbourhood function. We will look next at how to define the new neighbourhood functions that we will denote $\mathcal{N}_i \boxtimes \mathcal{E}_i$, defining them eventually in Definition 3.2. Once we have defined this, we will be able to write $(W, \mathcal{N}_i, \xi)_{i \in N} \boxtimes (\Sigma, \mathcal{E}_i, \text{PRE})_{i \in N}$ for:

$$(W \otimes \Sigma, \mathcal{N}_i \boxtimes \mathcal{E}_i, \xi \otimes \Sigma).$$

Recall that the interpretation of X being a neighbourhood of u for i in a state model is that i has the information that X. Now, when do we want to say that, at (u, e), i has the information that X?

Let us consider first of all the simpler case where there would be no pre-conditions, or (equivalently) where $\text{PRE}(e) = \top$ for all $e \in \Sigma$. Then we will add preconditions back once the simpler case is clear. Firstly, note that if X is not a *rectangular* subset, i.e. is not of the form $F \times E$ with $F \subseteq W$ and $E \subseteq \Sigma$, then i can only have this information on the basis of some other 'rectangular' information. The reason is that since we are in a hypothetical situation where all actions are without preconditions, there can be no 'correlation' between the states and actions. That is, if i has information that entails that (v, c) is not the actual state, it is because i has information that eliminates v or that eliminates c. (To reiterate: a 'correlation' of a kind does reappear when we move back to the situation we ultimately want to consider, in which certain actions are ruled out by certain states on the basis of preconditions.)

So for now we will consider just rectangular subsets, though since we are considering monotonic models, we will anyway want to close for supersets, so players will be able to have information that is not 'rectangular', but only on the basis of some piece of rectangular information. Take then some $X \in W \times \Sigma$, i.e. such that $X = F \times E$, where $F \subseteq W$ and $E \subseteq \Sigma$. When will i have information that X? It is only when X had the information that F, and 'received' (through whatever action just occurred) the information that E.

If i did not have the information that F, then she has learnt something 'miraculously', i.e. without any informational reason. Similarly, if i has not received the information that E could have occurred, then she has no basis for the information that $F \times E$. For the converse direction: suppose that i had the information that F, and received the information E. Then she has (this time legitimately!) learnt the information that $F \times E$.

There is a conceptual objection of a kind to be raised here: since in monotonic neighbourhood models players are unable to 'put together' their pieces of information, it might be said that while the player receives the information that E, she is unable to put this together with her information that F. We can only say that we assume a kind of perfection in the information-receiving capacities of the player that might be lacking in her information-storage capacities. Nonetheless, we do *not* consider it an uninteresting topic to look for generalisations of product update for agents who receive information in an imperfect way in the same sense that neighbourhood models might

be said to represent agents who process their existing information imperfectly, by not putting together all the pieces of information that they have.

Still, we say that in the precondition-less case, i has the information $F \times E$ at (u, e) just if i had the information that F at u, and received the information that E. And for other (non-rectangular) subsets X, i has the information that X just if i has some rectangular information $F \times E \subseteq X$. That is, i's neighbourhoods at (u, e) would be the monotonic closure of $\mathcal{N}_i(u) \times \mathcal{E}_i(e)$, i.e.

$$\{X \subseteq W \times \Sigma \mid \exists Y \in \mathcal{N}_i(u) \times \mathcal{E}_i(u) : Y \subseteq X\}.$$

Clearly this definition will yield a monotonic neighbourhood model. However, it does not take the preconditions into account. Luckily, while reasoning about the product update was simplified by removing them, still adding them back in is entirely straightforward: we simply relativise the neighbourhood functions to the smaller space! Thus we are left with the following definition:

Definition 3.2. The effect of a neighbourhood action on a neighbourhood function, denoted $\mathcal{N}_i \boxtimes \mathcal{E}_i$, is the following:

$$(\mathcal{N}_i \boxtimes \mathcal{E}_i)(u.e) = \Big\{ X \subseteq W \otimes \Sigma \Big| \exists Y \in \mathcal{N}_i(u) \times \mathcal{E}_i(u) : Y \cap (W \otimes \Sigma) \subseteq X \Big\}.$$

Now, Fact 3.2 states that $\boxtimes$ is properly speaking a 'generalisation' of $\otimes$, in the sense that if both neighbourhood functions are relational (closed for arbitrary intersections), then $\boxtimes$ yields the same result as $\otimes$.

Fact 3.2. *If each $\mathcal{N}_i$ and $\mathcal{E}_i$ are equivalent respectively to the relations $\mathcal{R}_i$ and $\rightarrowtriangle_i$, then $(W, \mathcal{N}_i, \xi)_{i \in N} \boxtimes (\Sigma, \mathcal{E}_i, \text{PRE})_{i \in N} = (W, R_i, \xi)_{i \in N} \otimes (\Sigma, \rightarrowtriangle_i, \text{PRE})_{i \in N}$.*

Again we can define the language $\mathcal{L}_{\boxtimes}$ as we did above for $\mathcal{L}_{\otimes}$, by adding to $\mathcal{L}$ operators of the form $\langle \mathcal{A}, e \rangle$. The next question to ask is whether we can give reduction axioms, to reduce the language $\mathcal{L}_{N,O,\kappa,\boxtimes}$ to the language $\mathcal{L}_{N,O,\kappa}$. The affirmative answer is given in Theorem 3.1.

Theorem 3.1. *For any formula in $\mathcal{L}_{N,O,\kappa,\boxtimes}$, there is a formula in $\mathcal{L}_{N,O,\kappa}$ that is equivalent over monotonic neighbourhood models.*

Proof. The proof involves giving a validity-preserving translation; the key step is to show that the following reduction axiom is valid for neighbourhood product update:

$$\langle \mathcal{A}, e \rangle \Box_i \varphi \equiv \text{PRE}(e) \wedge \bigvee_{E \in \mathcal{E}(e)} \Box_i \bigwedge_{d \in E} [d]\varphi$$

To see that validity, take any monotonic neighbourhood state model $\mathcal{M} = (W, \mathcal{N}_i, \xi)_{i \in N}$ and monotonic neighbourhood action model $\mathcal{A} = (\Sigma, \mathcal{E}_i, \text{PRE})_{i \in N}$. Then we have the

following chain of equivalences:

$$
\begin{aligned}
\mathcal{M} \boxtimes \mathcal{A}, (u,e) \Vdash \Box_i \varphi \quad &\Leftrightarrow \quad [\![\varphi]\!]_{\mathcal{M}\boxtimes\mathcal{A}} \in (\mathcal{N}_i \boxtimes \mathcal{E}_i)(u,e) \\
&\Leftrightarrow \quad \{(w,d) \in W \otimes \Sigma \mid \mathcal{M} \boxtimes \mathcal{A}, (w,d) \Vdash \varphi\} \in (\mathcal{N}_i \boxtimes \mathcal{E}_i)(u,e) \\
&\Leftrightarrow \quad \{(w,d) \in W \times \Sigma \mid \mathcal{M}, w \Vdash \langle \mathcal{A}, d\rangle \varphi\} \in (\mathcal{N}_i \boxtimes \mathcal{E}_i)(u,e) \\
&\Leftrightarrow \quad \exists Y \in \mathcal{N}_i(u) \times \mathcal{E}_i(e) : \\
&\qquad Y \cap (W \otimes \Sigma) \subseteq \{(w,d) \in W \times \Sigma \mid \mathcal{M}, w \Vdash \langle \mathcal{A}, d\rangle \varphi\} \\
&\Leftrightarrow \quad \exists Y \in \mathcal{N}_i(u) \times \mathcal{E}_i(e) : Y \subseteq ((W \times \Sigma) - (W \otimes \Sigma)) \cup \\
&\qquad \{(w,d) \in W \times \Sigma \mid \mathcal{M}, w \Vdash \langle \mathcal{A}, d\rangle \varphi\} \\
&\Leftrightarrow \quad \exists Y \in \mathcal{N}_i(u) \times \mathcal{E}_i(e) : \\
&\qquad Y \subseteq \{(w,d) \in W \times \Sigma \mid \mathcal{M}, w \Vdash [\mathcal{A}, d]\varphi\} \\
&\Leftrightarrow \quad \exists E \in \mathcal{E}_i(e)\, \exists F \in \mathcal{N}_i(u) : \\
&\qquad F \times E \subseteq \{(w,d) \in W \times \Sigma \mid \mathcal{M}, w \Vdash [\mathcal{A}, d]\varphi\} \\
&\Leftrightarrow \quad \exists E \in \mathcal{E}_i(e)\, \exists F \in \mathcal{N}_i(u) : \forall w \in F\, \forall d \in E,\, \mathcal{M}, w \Vdash [\mathcal{A}, d]\varphi \\
&\Leftrightarrow \quad \exists E \in \mathcal{E}_i(e) \exists F \in \mathcal{N}_i(u) : \forall w \in F,\, \mathcal{M}, w \Vdash \bigwedge_{d\in E} [\mathcal{A}, d]\varphi \\
&\Leftrightarrow \quad \exists E \in \mathcal{E}_i(e) \exists F \in \mathcal{N}_i(u) : F \subseteq \Big[\!\Big[\bigwedge_{d\in E} [\mathcal{A}, d]\varphi \Big]\!\Big]_{\mathcal{M}} \\
&\Leftrightarrow \quad \exists E \in \mathcal{E}_i(e) : \Big[\!\Big[\bigwedge_{d\in E} [\mathcal{A}, d]\varphi \Big]\!\Big]_{\mathcal{M}} \in \mathcal{N}_i(u) \;\; [\because \mathcal{N}_i \text{ monotonic}] \\
&\Leftrightarrow \quad \exists E \in \mathcal{E}_i(e) : \mathcal{M}, u \Vdash \Box_i \bigwedge_{d\in E} [\mathcal{A}, d]\varphi \\
&\Leftrightarrow \quad \mathcal{M}, u \Vdash \bigvee_{E\in\mathcal{E}(e)} \Box_i \bigwedge_{d\in E} [\mathcal{A}, d]\varphi
\end{aligned}
$$

Now we can substitute the first term for the last in the definition of the semantics of $\langle \mathcal{A}, d\rangle$, thus obtaining

$$
\begin{aligned}
\mathcal{M}, u \Vdash \langle \mathcal{A}, e\rangle \Box_i \varphi \quad &\Leftrightarrow \quad \mathcal{M}, u \Vdash \textsc{pre}(e) \;\&\; \mathcal{M}, u \Vdash \bigvee_{E\in\mathcal{E}(e)} \Box_i \bigwedge_{d\in E} [\mathcal{A}, d]\varphi \\
&\Leftrightarrow \quad \mathcal{M}, u \Vdash \textsc{pre}(e) \wedge \bigvee_{E\in\mathcal{E}(e)} \Box_i \bigwedge_{d\in E} [\mathcal{A}, d]\varphi
\end{aligned}
$$

■

Transfinite information addition

We now introduce a new kind of epistemic action: another variation on public announcement, that we think of as a kind of 'information addition'. We will specify the

semantics of this operator with respect to monotonic neighbourhood models (and so relational models) and topological models, where the definition will have to be changed slightly to ensure that the resulting model is topological. We also give reductions axioms for the basic modal language with a modality $\langle +\varphi \rangle$ corresponding to this operator, and show them to be valid with respect to all three types of semantics, and which can therefore be used to prove completeness with respect to all three semantics. We also show that this kind of announcement is what is required in order to generate the model described in Theorem 1.5, in which, as we indicated in Section 1.4, players' "ignorance about ignorance" allows for a model in which α-level mutual belief of rationality is equivalent to $1 + \alpha$ rounds of elimination of non-optimal strategies. This new operator adds information but in a manner that also creates 'ignorance' in the sense that for example it will transform an $S5$ model into an $S4$ model, so that players might initially be negatively introspective, i.e. they do not ignore any of their ignorance, but that after the announcement they will no longer be negatively introspective.

Recall that the elements of a player's neighbourhood of a particular model are supposed to represent the *pieces of information* possessed by the player in the relevant state. Then if we are to add the information that A, we simply add the event ('fact') A to the neighbourhood(s). That is in effect what is done in non-eliminative announcement: $\mathcal{N}_i{¡}A(u) = \{U \cap A \mid U \in \mathcal{N}_i(u)\}$. The present kind of public announcement that we consider is very much like this, except that it is in some sense a *conditional* announcement: it is made and believed just if it is true. So the action that we are considering works as follows:

$$\mathcal{N}_i + A(u) = \begin{cases} \{U \cap A \mid U \in \mathcal{N}_i(u)\} & \text{if } u \in A \\ \mathcal{N}_i(u) & \text{otherwise.} \end{cases}$$

We might therefore be tempted to define the model obtained by adding the information that A to $\mathcal{M}$, where A is an event in $\mathcal{M}$, as the same model $\mathcal{M}$ with the neighbourhoods appropriately substituted, so that in the new model i's neighbourhood function is given by $\mathcal{N}_i + A$. However, we want that if we are only considering relational, monotonic, or topological models then the resulting model is a relational, monotonic, or topological model.

Fact 3.3. *The following are immediate:*

1. *If $\mathcal{N}_i$ is relational then $\mathcal{N}_i + A$ is relational.*
2. *If $\mathcal{N}_i$ is monotonic then $\mathcal{N}_i + A$ is monotonic.*

Fact 3.3 means that we can safely set the semantic operation $+$ on models so that $\mathcal{M} + A$ is the same as $\mathcal{M}$ except that $\mathcal{N}_i$ is substituted with $\mathcal{N}_i + A$.

However, the topological case is not quite as immediate. There the analogous construction would be to turn τ into $\tau \cup \{U \cap A \mid U \in \tau\}$. Write $\tau \star A$ for $\tau \cup \{U \cap A \mid U \in \tau\}$; it is easy to see that τ being a topology does not guarantee that $\tau \star A$ be a topology.

Fact 3.4. *There is a topology τ over W and an event $A \subseteq W$ such that $\tau \star A$ is not a topology.*

Proof. Set $W = \{a, b, c, d\}$ and $\tau = \{\emptyset, \{a, b\}, \{c, d\}, W\}$, which is a topology over W. Then if $A = \{a, c\}$, $\{U \cap A \mid U \in \tau\}$ is not a topology, since for example it should contain $\{a, b, c\}$. ■

Nonetheless, notice that in the example in the proof of Fact 3.4, $\{U \cap A \mid U \in \tau\}$ is a *basis*. This is not a coincidence: Fact 3.5 shows that it holds in general, and indeed that $\tau \star A$ generates a topology (that is of course a refinement of τ) by taking just finite unions:

Fact 3.5. *If τ is a topology then $\{X \cup Y \mid X, Y \in \tau \star A\}$ is a topology.*

Proof. The proof is easy but we rehearse it nevertheless:

First, we show that $\tau \star A$ is closed for finite intersections: Take $X, Y \in \tau \star A$. If $X, Y \in \tau$ then we are done since τ is a topology. Otherwise, without loss of generality, we have either (1) $X \in \tau$ and $Y = U \cap A$ with $U \in \tau$, or (2) $X = B \cap A$ and $Y = U \cap A$ with $U, B \in \tau$. In case (1), $X \cap U \in \tau$, in which case $X \cap Y = X \cap U \cap A \in \tau \star A$. Similarly, in case (2), $U \cap B \in \tau$, so $X \cap Y = B \cap U \cap A \in \tau \star A$.

Next, we show that $\tau \star A$ is 'almost closed' for arbitrary unions, in the sense that for any $\Xi \subseteq \tau \star A$, there are $X, Y \in \tau \star A$ such that $X \cup Y = \bigcup \Xi$. To show this, take any $\Xi \subseteq \tau \star A$. Then we know that $\Xi = P \cup Q$ where $P \subseteq \tau$ and there is some $R \subseteq \tau$ such that $Q = \{U \cap A \mid U \subseteq R\}$. Let $X = \bigcup P$, $Y = \bigcup Q$ and $Z = \bigcup R$. Then since τ is closed for arbitrary unions, we know that $X \in \tau \subseteq \tau \star A$, and $Z \in \tau$, so that $Z \cap A \in \tau \star A$. But $Z \cap A = Y$, so we are done. ■

Fact 3.5 means that we can safely define $\tau_i + A$ as follows:

$$\tau_i + A = \{X \cup Y \mid X, Y \in \tau_i \star A\}.$$

Then given a topological model $\mathcal{M}$, we define the model $\mathcal{M} + A$ to be the same model except that each player i's topology is now given by $\tau_i + A$ rather than τ_i.

In all cases, we write $+\varphi$, where φ is a formula, to mean $\mathcal{M} + [\![\varphi]\!]_{\mathcal{M}}$.

On the syntactic level, we can enrich the language with a binary modality $\langle +\varphi \rangle \psi$. As with other public announcement operators, this is a *model-changing* modality, so has the following semantics:

$$\mathcal{M}, u \vDash \langle +\varphi \rangle \psi \text{ iff } \mathcal{M} + \varphi, u \vDash \psi.$$

We would like to give reduction axioms for this modality, showing that the language with the modality has exactly the same expressivity as the original language. This again also means that given an axiomatisation of the rest of the language, we can obtain completeness for the language enriched with the $\langle + \rangle$ modality. Proposition 3.6 gives the correct reduction axioms for this modality, that again lead to a truth-preserving translation.

Proposition 3.6. *The following reduction axiom is valid on monotonic neighbourhood models and on topological models.*

$$\langle +A\rangle\Diamond\varphi \quad \equiv \quad \Diamond\langle +A\rangle\varphi \wedge (A \to \Diamond(A \wedge \langle +A\rangle\varphi)$$

And, as in Proposition 3.7, the reduction axioms from Proposition 3.4 yield valid analogues for the non-epistemic parts of the language.

As we noted in Fact 3.3, if $\mathcal{M}$ is relational then $\mathcal{M}+A$ will be relational. However, the same does not hold in the limit. That is, it is possible to keep on announcing a fact and thereby transform an intersection-closed neighbourhood model into a model that is no longer intersection-closed, i.e. into a model that is no longer equivalent to a relational model. A simple example can be used to illustrate this phenomenon: Let $N = \{a\}$, $T_aW = \mathbb{N} \cup \{\omega_0\}$; set $\mathcal{N}_a(u) = W$ for all $u \in W$, and $\xi(u) = u$. Then clearly $\mathcal{M} = (W, \mathcal{N}_a, \xi)$ is an intersection-closed model, equivalent to the relational model $(W, W \times W, \xi)$.

Consider the following model, generated by an infinite sequence of conditional announcements

$$\mathcal{M} + \neg\mathbf{0} + \neg\mathbf{1} + \ldots + \neg\mathbf{m} + \ldots$$

Although at each stage of building this model it is relational, in the limit it is no longer relational. Let $\alpha^{>} = \{n \in \mathbb{N} \cup \{\omega_0\} \mid n > \alpha\}$, and then define $\mathcal{N}^\alpha$ recursively:

$$\begin{array}{rcl} \mathcal{N}^0 & = & \mathcal{N}_a \\ \mathcal{N}^{\alpha+1} & = & \mathcal{N}^\alpha + \alpha^{>} \\ \mathcal{N}^\lambda & = & \bigcup\limits_{\alpha<\lambda} \mathcal{N}^\alpha. \end{array}$$

Then by Fact 3.3, for every $m \in \mathbb{N}$, $\mathcal{N}^m$ is relational. However, there will also be a point, in the transfinite case, where we are in a model that is not intersection-closed. Notably, we have the following:

$$\{m^{>} \mid m \in \mathbb{N}\} = \mathcal{N}^{\omega_0}(\omega_0),$$

Yet we also have

$$\bigcap\{m^{>} \mid m \in \mathbb{N}\} = \{\omega_0\},$$

but clearly $\{\omega_0\} \notin \mathcal{N}^{\omega_0}(\omega_0)$.

Remember that we needed to move to non-relational models in order to give a correct foundation for α rounds of iteration of non-optimal strategies for transfinite α (cf. Theorem 1.5), and the initial model for a game, as defined in Section 3.2, will be relational. So this fact that the $+$ announcement operator can, with enough iterations, turn a 'relational' neighbourhood model into a non-relational model will be useful when we look, as we do in the next Section, at what information flows can create the models like those in Theorem 1.5.

Definition 3.3. We will write $\mathcal{M} +^{\alpha} \varphi$ to mean the model generated by repeatedly applying $+\varphi$, α times, where α is an arbitrary ordinal. For the limit case, we use essentially the same definition as the one we just gave: suppose that

$$\forall \beta < \lambda,\ \mathcal{M} +^{\beta} \varphi = (W, \mathcal{N}_i^{\beta}, \xi)_{i \in N}$$

is defined; then set

$$\mathcal{M} +^{\lambda} \varphi \ = \ (W, \bigcup_{\beta < \lambda} \mathcal{N}_i^{\beta}, \xi)_{i \in N}.$$

This definition allows us to consider the effects of transfinite announcements.

3.2 Epistemic actions on games

Just as the optimality operators considered in Chapter 1 in effect reduce the game, so do public announcements reduce the game model. Starting with an initial model $\mathfrak{I}_G$ of a game G, if one makes an announcement that has the effect of saying that the players will only play according to the subgame S, then one obtains a model $\mathcal{M}_{G_S}$ of the game G_S that has the strategies in S, with the preferences over them being the same as the preferences over them in the original game G. It seems reasonable to ask that that model $\mathcal{M}_{G_S}$ be the initial model $\mathfrak{I}_{G_S}$ of the smaller game G_S.

This idea is taken direction from [Benthem, 2007b]. We will not touch upon the main technical contributions of that paper however. So all we do is use it as a springboard for discussion, and so although we mention it frequently, what we say should certainly not be taken as in any sense summarising it.

There are many ways to interpret what a public announcement of optimality or rationality in the game might be. Let us first consider the interpretation suggested in [Apt and Zvesper, 2007], that public announcements can be made by players, to the effect that they will not play such-and-such strategies. Then each public announcement is associated with a player i, and can only eliminate strategies of player i. We will call these public announcements *individual public announcements*. Thus if the language can express strategies, an individual public announcement by player i could, syntactically speaking, be of the form $[!\neg \bigwedge S_i]$, where $S_i \subseteq T_i$.

However, as in the rest of this Chapter, we will be interested in generating restrictions in an homogeneous way. That is to say, we want to consider the case in which the public announcements are *syntactically* the same between different rounds. In particular, the natural choices for our immediate concerns will be that each player announces her own rationality, or that she is playing optimally. Therefore rather than considering only announcements of the form $[!\neg \bigwedge S_i]$, we will look at the more general class of announcements of the form $[!\varphi]$ where, *in the model being considered*, φ defines a subset of the model which is the interpretation of some sentence $\bigvee S_i$, where $S_i \subseteq T_i$. (More strictly we should say, 'would be equivalent were that last sentence in the language', for we will not assume that it is in the language.)

Definition 3.4. An ***individual public announcement*** by i in the the model $\mathcal{M}$ is an announcement $[!\varphi]$ where φ is an arbitrary formula and, for some $S_i \subseteq T_i$, $[\![\varphi]\!]_{\mathcal{M}} = [\![\bigwedge S_i]\!]_{\mathcal{M}}$. To put it otherwise: $\xi([\![\varphi]\!]_{\mathcal{M}}) = S_i \times T_{-i}$ for some $S_i \subseteq T_i$.

The relational model that is taken in [Benthem, 2007b] to represent the initial situation before any announcements have taken place uses the strategy profiles as the state space. In the model, at every state each player is taken to be correct about her own strategy, and to have no belief about what strategy the other players will play, indeed in some sense the *only* information each player has is about her own strategy: she considers possible all states where she plays the same strategy.[2] So given some game $G = (T, <)$, the model would be the relational model $\mathfrak{I}_G = (W, R_i, \varepsilon)_{i \in N}$, with ε the identity function and $R_i(s) = \{s_i\} \times T_{-i}$.

This epistemic relation is precisely the strategy relation that we considered including in the semantics of a language in the previous section, to interpret the modality $[i_s]$.

In the context of such a model $\mathcal{M}$, it is certainly clear that players can legitimately make a large number of individual public announcements, since they do indeed correctly believe what strategy they will play. Thus each player i can 'honestly' announce, at u, any individual public announcement φ such that $\xi_i(u) \in \xi_i([\![\varphi]\!]_{\mathcal{M}})$.

Notice then that in case the players are playing according to iterated elimination of non-optimal strategies then they can each simply announce this, and we will have a model in which the players each believe they will all play according to the iterated elimination of non-optimal strategies. But the much more interesting line pursued in [Benthem, 2007b] involves studying repeated announcements that the players are *rational*.[3]

It is important to note that the particular interpretation we are considering here is not necessarily that intended by [Benthem, 2007b]. It is not entirely clear from that paper what interpretation should be given to the announcements; they are studied rather in the spirit of connecting different research fields, and illustrating the dynamic nature of contemporary mathematical and philosophical logic.

Furthermore, let us note that it is unclear what situation this model is intended to represent, since it seems that players should have some belief about the strategy of the other players before deciding on their own strategy. Yet in this model the players have a determinate belief about what they will do – each has decided her own strategy – apparently without any information about what the other players will do. Thus it would

[2] Since that fact is itself commonly believed, the model does not really represent *total* ignorance on the part of the players; for example each player believes that the other players are correct about what they will play, and so on.

[3] Two versions of rationality are considered in [Benthem, 2007b]: "weak rationality" and "strong rationality", corresponding to avoiding strictly dominated strategies and avoiding never-best responses respectively. The definability of the resulting outcomes in an inflationary fixpoint calculus is considered, and observations about monotonicity are given more formal force by observing a common syntactic form in terms of "existential positive" formulae.

be difficult to wrap an intuitive interpretation around the mathematical description proposed.

As we have suggested already, the models that we will be interested in represent our interpretation of the one-shot interaction situation. We therefore define the *initial model* of a game $\mathfrak{I}_G$ of a game G as consisting of one state for each strategy profile, with complete uncertainty for all players concerning the states

Definition 3.5. If $G = (T, <)$, then the ***initial model*** of G is $\mathfrak{I}_G = (T, \mathcal{N}_i, \xi)_{i \in N}$, with $\mathcal{N}_i = \{T\}$ for each player $i \in N$.

(Notice that this initial model's neighbourhoods are trivially monotonic for each player, and of course each public announcement cannot break the monotonicity. So the syntactic analysis from the previous Section would apply here unproblematically.)

These models are very simple: a far cry from the '(assumption-)complete' or 'universal' models mentioned in Chapter 2. Nonetheless, we think they are faithful to the one-shot situation as it is described: players are presented with (or perhaps 'confronted with') the game situation, and the situation is assumed to be common knowledge. So it is not the case that the players do not know *anything*; in particular they know the game, and the epistemic situation of themselves and the other players. Ideally a full informational account of game theory might start with some much more general initial situation and describe dynamically the process of acquisition of the game situation as it is described by our initial models. However, we think that our models are intuitively plausible and that they do justice to the one-shot interpretation of strategic games as we have described it.

Notice in particular that in these models players do not have any beliefs regarding their own strategy. This is in contrast to the situation as it is described in [Benthem, 2007b]. In our attempt to model a deliberative process, we have players choosing their strategies *after* reasoning about each other. So rationally they eliminate choices until they are unable to continue to do so, and *then* make a choice. They might make this choice realising that they and the others have several possible rational choices.

The process of elimination itself we describe as a ***private but common*** process, since the idea behind it is that all players must suppose that the other players are performing the same process, so that the model is updated in the same way. Thus reasoning is done *privately*, since this is still meant to represent the one-shot situation, but *commonly*, since all players are in some sense in the same situation.

Therefore we suggest that epistemic models in which players have settled on their strategies, after this process of rational, private but common, deliberation, should be arrived at via those two processes: First, the private but common deliberation, described by the public announcements. Then a *decision* by each player, to choose one of the strategies she has left. Now these decisions also have a 'private but common' character to them: the detail of them is private, but the players are commonly aware that each other player is making some decision. The players all making a decision, and being aware that the others have all made their decision, is given by taking a model $\mathcal{M} = (W, \mathcal{N}_i, \xi)_{i \in N}$ and returning the updated model $\mathcal{M}\mathrm{D}$.

Definition 3.6. $\mathcal{M}\mathrm{D} = (W, \mathcal{N}_i\mathrm{D}, \xi)_{i\in N}$:

$$\mathcal{N}_i\mathrm{D}(u) = \{E \subseteq W \mid \exists A \in \mathcal{N}_i(u) :\ A \cap \xi_i^{-1}(\xi_i(u)) \subseteq E\}.$$

What this operation does is to ensure that each player is correct about their own strategy choice, and that fact becomes commonly believed among the players. In the case of a relational model, it can be written as follows:

$$R_i\mathrm{D}(u) = R_i(u) \cap \xi_i^{-1}(\xi_i(u)).$$

Now, as in the case of public announcements, we can talk about the new model in the old model. That is: we can give 'reduction axioms' for a modality $\langle \mathrm{D}\rangle$ that allow us to provide a truth-preserving translation between $\mathcal{L}_{N,O,\kappa,!}$ and $\mathcal{L}_{N,O,\kappa}$.

Proposition 3.7. *The following reduction axiom is valid:*

$$\langle \mathrm{D}\rangle \Diamond_i \varphi \ \equiv\ \bigwedge_{i\in N}\Big(\bigwedge_{s_i\in T_i} \Big(s_i \to \Diamond_i(s_i \wedge \langle \mathrm{D}\rangle\varphi)\Big)\Big)$$

Furthermore, because this operation also only alters the information of players in a model, all the reduction axioms for $\langle ¡ \rangle$, other than that for $\Diamond_i$, remain valid when we replace $\langle ¡ \rangle$ by $\langle \mathrm{D}\rangle$.

We can use non-eliminative announcements, followed by each player making their decision, to generate a model that is somewhat like that given in Theorem 1.2. That Theorem stated the existence of a model in which players have common true belief in rationality just if they all play strategies that survive the iterated elimination of non-optimal strategies.

So take some game G, and its initial model, as defined above $\mathfrak{I}_G = (T, \mathcal{N}_i, \varepsilon)_{i\in N}$. Then, in [Benthem, 2007b], we can repeatedly announce eliminatively that each player plays optimally, or, *equivalently*, that she plays rationally. This will yield a model in which players have common belief of rationality. In this model, the 'rationality' $\mathbf{r}_i$ of each player i is *the same thing as* i playing according to the iterated elimination of non-optimal strategies. In order to obtain a model in which players are, in addition, correct about their strategies, and in which this fact is (commonly) believed, we apply the D operator that we introduced above:

$$\mathfrak{I}_G \underbrace{!\bigcirc\top \ldots !\bigcirc\top}_{\alpha_G \text{ times}}\mathrm{D} \ =\ \mathfrak{I}_G \underbrace{!\mathbf{r} \ldots !\mathbf{r}}_{\alpha_G \text{ times}}\mathrm{D}$$

However, this model is in other ways different from the model constructed in Theorem 1.5, since there are no states in this model where the players do not play rationally, whereas for most games there *are* in states in the model given by that construction, in which players do not play rationally. (By 'most games', we mean all those in which the optimality operator in question is ***non-tautological***, in the sense that it eliminated at least one strategy of at least one player.)

To get closer to that model, we could instead apply *non-eliminative* public announcements. Fact 3.6 shows that in this case we must use announcements of *rationality* and not just optimality.

Fact 3.6. *Non-eliminatively announcing rationality more than once has no more effect than doing so only once:*

$$\mathcal{M}_{¡} \bigcirc \top = \mathcal{M}_{¡} \bigcirc \top_{¡} \bigcirc \top$$

So we could get closer to seeing how the model from Theorem 1.5 might come about by considering the following process:

$$\mathfrak{I}_G \underbrace{¡\mathbf{r} \ldots ¡\mathbf{r}}_{\alpha_G \text{ times}} \mathrm{D}$$

However, this model is also different from that given in Theorem 1.5, since in the states where players do not play rationally, at any stage of the process, they acquire inconsistent beliefs, i.e. in those states, each player's neighbourhood will contain the empty set $\emptyset$. However, in those states that are left in the outcome, we do indeed have common belief that players are rational, and are correct about their own strategies. It *is* possible to define another non-eliminative announcement that does not have this effect. We could specify another ad-hoc operation and give a reduction axiom for it, but we can also define a DEL action model to achieve the same effect.

So, as we will now see, there is an action model that can be used to generate, given an initial model $\mathfrak{I}_G$ for the game G, as above, a model similar to that in Theorem 1.2. That model was an S5 (partitional) model, and in it common belief of rationality was

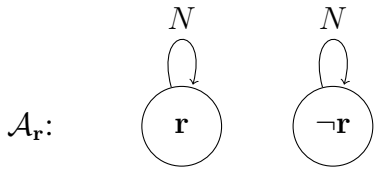


Figure 3.4: An action model that gives a dynamic counterpart to Theorem 1.2. Here there are two events, one with the precondition $\mathbf{r}$, the other with the precondition $\neg\mathbf{r}$; and all players can tell which event is occurring.

equivalent to the iterated elimination of non-optimal strategies. In order to generate it, we simply give, as depicted in figure an action model that is the disjoint union of two different public announcements.

The action depicted in Figure 3.4 is just an announcement '*whether*' the players are rational. That is, at states where the players are rational, it functions just like an announcement that they are rational; at states where not all players are rational, it is an announcement that not all players are rational.

We can see the action working on an initial model of some game G, which we draw in Figure 3.5. There we draw the (relational model of the) two-player game as a square. The optimality operators for each player are marked along the side, and the accessibility relation (which is the same for each player) is given by a dashed line indicating the *partition* induced by it (since the models are all S5). The model on the far left is the initial model $\mathfrak{I}_G$, and we successively apply the action model $\mathcal{A}_{\mathbf{r}}$ from Figure 3.4, to generate new models, with new accessibility relations. That model at the

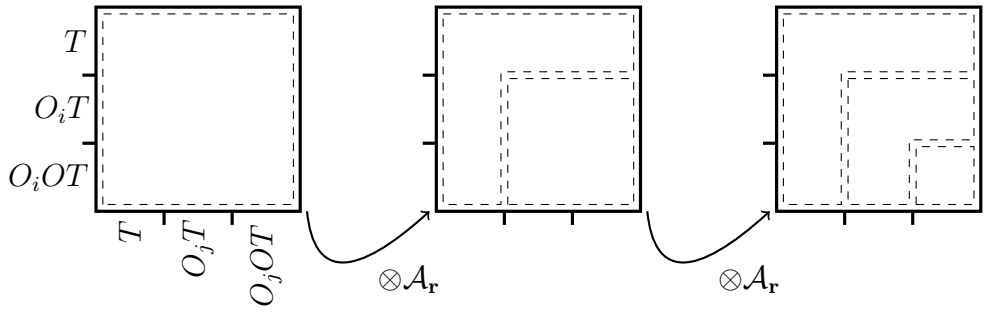


Figure 3.5: The announcement from Figure 3.4 being applied to an initial model. We depict a two-player game model by arranging the states into rows and columns according to the strategy choices of a row and a column player. Each application of $\mathcal{A}_{\mathbf{r}}$ refines the players' information (they have the same information) which is a partition denoted by the dashed lines.

far right represents the situation where the players have been successively 'informed' whether or not they are all playing rationally. Subsequently, the players choose their strategy, i.e. we apply D to the model, as in Figure 3.6. (For illustration we suppose that each player has only three strategies.) Here where the two partitions are different we draw the column player's partition dashed as before, and the row player's partition as dotted lines.

Notice that this is *not* the same model as that given in Theorem 1.2, since here the players have more information than in that model. In that model, there were only two elements in each player's partition, which were the event that players play according to O^∞, and the event that they do not. This model is relatively simple to generate however; it is not clear what action would be *iterated*, in step with the algorithm of elimination, in order to generate a model like that in Theorem 1.2.

Let us now turn to the case of transfinite announcements. In the rest of this Section, we show that there is a single statement that, when '*conditionally*' announced (so in '$+\varphi$' sense) α times, generates a model where, like that in Theorem 1.5, for all $\beta < \alpha$, β-level belief in rationality is *equivalent* to $1 + \beta$ rounds of elimination of non-optimal strategies.

We will illustrate this in the 2-player case; the statement however is already a little more complicated than just 'both players are rational'. Let us build it up step by step.

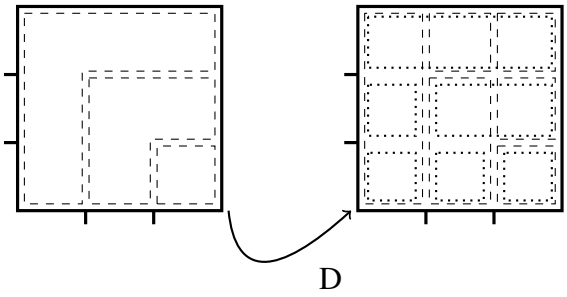


Figure 3.6: The situation once each player has chosen a strategy. Here the two players get different information, since each is aware of his own choice but ignorant of the opponent's.

First, consider the following formula scheme, where j is the player who is not i:

$$\varphi_{S_i} := (\Diamond_i \Box_j S_i \wedge \bigcirc_i \bigcirc_j S_i) \rightarrow \bigcirc_j S_i$$

What this says, if you are player i, can be paraphrased as: 'If you think it's plausible that your opponent is rational and believes that you will play according to S_i, and you're playing optimally against that eventuality, then you are right to do so.'

Now what we want is to announce that this holds no matter what the strategy set S_i, and for both players. So let $\varphi_{\mathbf{cr}} := \bigwedge_{i \in N} \bigwedge_{S_i \subseteq T_i}$. This sentence is certainly not as straightforward as just announcing the rationality of both players, but Proposition 3.8 states that the model generated by α rounds of 'conditionally' announcing it, according to Definition 3.3, starting from the initial model of some game G, yields a model satisfying the condition of Theorem 1.5: that $O_G^{1+\alpha} = \xi([\![\mathbf{r} \wedge \Box^\alpha \mathbf{r}]\!])$.

Proposition 3.8. *Let $\mathcal{M}$ denote $(\mathfrak{I}_G +^\alpha \varphi_{\mathbf{cr}})$D. Then $\forall \beta \leq \alpha$ we have the following equivalence:*

$$\xi([\![\mathbf{r} \wedge \Box^\beta]\!]_{\mathcal{M}}) = O_G^{1+\beta}.$$

3.3 Belief revision and lexicographic rationality

In this Section we define a richer kind of belief model than that used so far, which can be used for reasoning about *conditional beliefs*, and allows for what [Benthem, 2007a] calls "soft information" flow. Partly this is groundwork for the next Chapter, where we will want models in which players can learn things, *in the course of playing a game* that contradict things they believed to be true, and so *revise* their information. However, we also show that conditional belief models are important for understanding the dynamic process of reasoning involved in the iterated application of *non-monotonic* optimality operators, like the elimination of weakly dominated strategies.

We will suggest that, during the decision phase, a player will choose precisely those strategies *from the original game* that are rational *given his information in whatever*

model has been arrived at by deliberation until the decision phase. So the kind of belief model that we end up with should be what we call a '***rational equilibrium of beliefs***', in which players recognise the possibilities that are rationally open to themselves and to each other, and, crucially, recognise the ***stability*** of the configuration of beliefs.

It happens that in the case of monotonic operators, the stability we are talking about is present in the resulting model. So in order to be clearer about what we mean, let us consider now a particular non-monotonic optimality notion: weak dominance (admissibility). An example should clarify what we mean, and serve to point the way towards the solution. In the game depicted in Figure 3.7, weak dominance would

	L	R
U	$0, 0$	$1, 0$
D	$1, 0$	$0, 1$

Figure 3.7: A strategic game where hard announcements of admissibility yield an unstable belief model.

eliminate first of all L for the column player b, because it is weakly dominated by R. Once L has been eliminated, D is then (strictly) dominated by U for the row player a, yielding the unique outcome (U, R). The crucial point about this simple example though is that *b's reason for not playing L is removed by one of its consequences*. That is, there is a sort of non-monotonicity in the reasoning. What that means is that, with respect to the restriction $\{(U, R)\}$, it is rational for player b to play L, since given that player a will (apparently) choose U, player b believes she is choosing between two equal alternatives.

The idea that now introduce of a *rational equilibrium of beliefs* is that the beliefs of the players concerning the strategies that might be played are *self-enforcing*; that if a player j believes that another player i's strategy s_i will not be played, then it is because the configuration of beliefs means that s_i would not be rational for i to play.

Definition 3.7. There is ***rational equilibrium*** at u in the relational model $(W, R_i, \xi)_{i \in N}$ iff :

$$\forall i \in N, \forall s_i \in T_i \Big(s_i \in O_i(R_i(u)) \Rightarrow \forall j \in N - \{i\}, (W, R_i, \xi), u \Vdash \Diamond_j s_i \Big).$$

We do not claim that Definition 3.7 is the final word on what a rational equilibrium in a relational model should be. Most notably, an obvious objection is that it might be better to stipulate that if j thinks i won't play s_i then it is because j *thinks* s_i would not be rational for i. However, we stick to the simple definition for present purposes since in all the examples we consider all beliefs are anyway common beliefs. A more in-depth study of these questions would not be difficult but might detract away from the our main point here.[4]

[4]Furthermore, note that an analogous definition could be given for the case of neighbourhood models.

Where G is the game in Figure 3.7, in the model $\mathfrak{I}_G!\mathbf{rD}$ arrived at by in effect announcing (eliminatively) that the players do not play weakly dominated strategies, there is one unique state left, that where players play (U, R). But this is *not* in rational equilibrium, for precisely the reason we gave above: b's reason for not playing L has been eliminated. I.e., it *would* be rational for b to play L, but a believes that b will not play L.

The way to overcome this problem is to remark that b can entertain the possibility that she might be mistaken in her belief. That is, since *given her beliefs* L and R appear equal, she should fall back on the possibility that she might be wrong: that a could, contrary to her information, play D.

So it looks like the solution might be to use non-eliminative announcements in combination with the universal modality, as a stricter form of the belief operator (that we might be tempted to call 'knowledge', since it exhausts every possibility in the model). Then we would say that a player plays rationally just if she plays optimally with respect to her beliefs *and that amongst those rational options* she only picks options that are optimal with respect to the rest of the model.

	L	R
W	$1, 0$	$-1, 0$
M	$0, 1$	$1, 0$
D	$0, 0$	$0, 1$

Figure 3.8: A strategic game that motivates using a finer-grained view of beliefs.

However, as the game in Figure 3.8 illustrates, we need a more fine-grained approach than that. In that example, (W, L) is the unique outcome of iterated admissibility. Notice now that the 'information' that this is the outcome does not leave the column player b with any reason to play according to the outcome, i.e. she might just as well decide to play R: this would still be rational. So among the options that she has that are optimal against W (both of them: L and R), we should then look at the options she has that are optimal with respect to all the possibilities in the model, i.e. against $\{W, M, D\}$. But in that case, R still has not become irrational for her! It is supported by the possibility that player a will choose D. Thus once again we do not have a rational equilibrium in the relevant model.

So we need a finer-grained approach: we want to say that player b considers it *most* likely that player a will play W; but believes that that if he doesn't then it's most likely that he will play M; and finally, considers the *least* likely option to be that he will play D.

This leads us to use the idea of a *plausibility* ordering over the states of the model, that we will now use to define beliefs *including conditional beliefs*. We therefore define *plausibility models*, which are still state-space based models (though along the lines we saw in Chapter 2, it would be possible to consider also type-space based models).

Definition 3.8. A ***total plausibility ordering*** is a total transitive reflexive relation, i.e. $\preccurlyeq_i$ is a total plausibility ordering iff :

- $\forall u, v \in W$, either $u \preccurlyeq_i v$ or $v \preccurlyeq_i u$.
- $\forall u, v, w \in W$, if $u \preccurlyeq_i v$ and $v \preccurlyeq_i w$, then $u \preccurlyeq_i w$.
- $\forall u \in W$, $u \preccurlyeq_i u$.

Definition 3.9. ***Plausibility models*** are of the form $(W, \preccurlyeq_i, \xi)_{i \in N}$, where W is a finite set of 'states', ξ is as before a function assigning to each state an outcome of the game, and each $\preccurlyeq_i \subseteq W \times W$ is a *plausibility ordering*.

(We restrict our attention here to finite models just to avoid questions of well-foundedness of the relation that would be entirely peripheral to our main concerns.)

We follow [Board, 2002; Benthem, 2007a; Baltag and Smets, 2006] in adopting plausibility orderings to represent *conditional beliefs*. So we henceforth consider languages with conditional belief operators:

$$\varphi ::= p \mid \neg\varphi \mid \varphi \wedge \varphi \mid \Box_i(\varphi|\varphi) \mid \dots$$

The conditional belief operator $\Box_i(\varphi|\psi)$ is supposed to mean something like 'i believes *conditionally on* ψ that φ'. More precisely, it will mean that in all i's most plausible ψ-states, φ holds. On a plausibility model $\mathcal{M} = (W, \preccurlyeq_i, \xi)_{i \in N}$, we interpret it as follows:

$$\mathcal{M}, u \Vdash \Box_i(\varphi|\psi) \quad \text{iff} \quad \text{MIN}_{\preccurlyeq_i}([\![\psi]\!]_{\mathcal{M}}) \subseteq [\![\varphi]\!]_{\mathcal{M}}.$$

Logics like this conditional belief logic are studied in [Segerberg, 1995; Board, 2002], cf. [Lewis, 1973].

Sometimes, when it leads to more elegant notation (notably in Chapter 4), we will write $\Box_i^{\varphi}\psi$ for $\Box_i(\psi|\varphi)$. We retrieve the unconditional belief operator $\Box_i\varphi$ by simply defining it as an abbreviation of $\Box_i(\varphi|\top)$.

In fact, plausibility models can be thought of as enrichments of positively and negatively introspective relational models. Unpacking the definition, we see that the unconditional belief modality has the following semantics:

$$\mathcal{M}, u \Vdash \Box_i\varphi \quad \text{iff} \quad \{v \in W \mid \forall w \in W,\, v \preccurlyeq_i w\} \subseteq [\![\varphi]\!]_{\mathcal{M}}.$$

We can talk about *common belief* $\Box^*$; (we could also introduce 'common conditional belief' $\Box^*(\varphi|\psi)$, but we will have no need for it). For every conditional belief model $(W, \preccurlyeq_i, \xi)_{i \in N}$ there is a relational model $(W, R_i, \xi)_{i \in N}$ defined by setting

$$R_i(u) = \text{MIN}_{\preccurlyeq_i}(W),$$

and this relational model will be equivalent to the conditional belief model for the language without the conditional belief modality.

It is not hard to see that beliefs in this relational model will be introspective and consistent, and equivalent to those of the conditional belief model. What is more, it is possible to show that the languages considered in the previous section cannot distinguish between on the one hand the class of relational models in which D, 4 and 5 hold, and on the other the class of plausibility models. That is: they have the same logic, as far as unconditional beliefs go.

So what exactly are these *un*conditional beliefs? Let us mention briefly the connection between conditional belief models and conditional probability systems. If, for each player i, we are given a *conditional probability system* à la Renyi [1955], or a *lexicographic probability system* [Blume *et al.*, 1991], over a set of states W we can define subjective conditional probabilities $\text{Prob}_i(F|E)$ even for events of zero probability. When W *is finite and the system is discrete* (i.e., $\text{Prob}_i(F|E)$ is defined for all non-empty events E), we can use this to define conditional belief operators for arbitrary events, by putting: $u \preccurlyeq_i v$ iff $\text{Prob}_i(\{u\}|\{u,v\}) \neq 0$. This will yield the following definition of the conditional belief operator:

$$[\![\Box_i(\varphi|\psi)]\!] = \{s \in S : \text{Prob}_i([\![\varphi]\!]|[\![\psi]\!]) = 1\}.$$

In the context of plausibility models, we will also be interested to define a notion of *knowledge*, the product of *hard information*. Let us therefore introduce a knowledge operator K_i into the language. We take the following axioms to be minimal requirements for a knowledge operator:

- $K_i\varphi \rightarrow \varphi$;
- $K_i\varphi \rightarrow \Box_i\varphi$.

These axioms should be valid if we accept that necessary conditions for you to *know* a proposition are firstly that it is true, and secondly that you believe it. (Of course, in any reasonable account of knowledge there will be more than just these minimal necessary conditions.) Then one way to ensure that these axioms hold would be to define each player i's knowledge modality K_i as the universal modality A. In that case, there will be no difference between the knowledge of player i and of player j, and so common knowledge, which we will write as K^* (the natural analogue to common belief) would be definable immediately also as A. We do not find this approach to be objectionable, especially in lieu of our strict interpretation of knowledge. However, we will note that there is another way, already present in [Board, 2002], of defining knowledge, in such a way that there can be propositions φ such that $K_i\varphi \wedge \neg K_j\varphi$ holds at some state. We will use this definition in Chapter 4.

Given a relation $\preccurlyeq_i \subseteq W \times W$, write W_i^u for the set of states that are $\preccurlyeq_i$-accessible from u, or from which u is $\preccurlyeq_i$-accessible:

$$W_i^u = \{v \in W \mid u \preccurlyeq_i v \text{ or } v \preccurlyeq_i u\}.$$

Definition 3.10. We say that the relation $\preccurlyeq_i$ is ***locally total*** if and only if:

- $\forall u \in W$, $\forall v \in W_i^u$, $W_i^v = W_i^u$.

Definition 3.10 can be paraphrased as saying that $\preccurlyeq_i$ is totally local just if the restriction of $\preccurlyeq_i$ to each W_i^u is total.

Definition 3.11. 1. A plausibility model in which the plausibility orderings are total is a ***pure plausibility model***.

2. A plausibility model in which the plausibility orderings are *locally total* is called an ***impure plausibility model***.

In Board's [2002] terminology, W_i^u is the set of states that are '*conceivable*' for i at u. It is a partition of the state space W and can be used to define the knowledge operator. We will assume the following definition of knowledge, that in *pure* plausibility models coincides with the definition of A the global modality:

$$\mathcal{M}, u \Vdash K_i\varphi \quad \text{iff} \quad W_i^u \subseteq [\![\varphi]\!]_{\mathcal{M}}.$$

$\Box_i(\varphi|\psi)$ does *not* mean that after learning that ψ, i will believe φ; rather it means that after learning ψ, i will believe that φ *was the case before the learning*. This is a subtle but important point: the conditional belief operators do not directly capture the dynamics of belief, but rather as van Benthem [2007a] puts it, they 'pre-encode' it. (We refer to [Benthem, 2007a; Baltag and Smets, 2008b] for more discussion.)

As explained in [Board, 2002, Section 3], we can use a binary plausibility ordering for each player, rather than a ternary ordering, indexed also by the actual state, (if and) only if we are willing to accept these strong introspection principles:

1. $\Box_i(\varphi|\psi) \rightarrow \Box_i(\Box_i(\varphi|\psi)|\chi)$
2. $\neg\, \Box_i\, (\varphi|\psi) \rightarrow \Box_i(\neg\, \Box_i\, (\varphi|\psi)|\chi)$

These two principles correspond respectively to a strong form of positive introspection and a strong form of negative introspection. The first entails 4, and the second 5, where we define the unconditional belief modality $\Box_i$ as above.

Although we accept that the resulting KD45 logic for the basic belief modality is potentially objectionable enough, and that the full introspection considered here might be considered worse, nonetheless this again is tangential to our main concerns, so for reasons of elegance (of notation if nothing else) we prefer to present the semantics with binary relations, and so assume full introspection of the players. (Recall from Chapter 1 that in type-space models, players are assumed to be fully introspective.)

The examples illustrated in Figures 3.7 and 3.8, and our discussion of them, point to a more fine-grained definition of rationality. Rather than only looking at what her *unconditional* beliefs are, a player should also take into account those situations that she would fall back on, if informed that her actual beliefs are false. Suppose that i cannot 'break a tie' between two strategies s_i and s_i' with respect to the game determined by her actual beliefs: by the $\preccurlyeq_i$-minimal states. Then i should, rationally, also

play optimally with respect to the 'next level' of beliefs. Let us formalise this notion of 'lexicographic rationality' (so called because players look first at their most likely states; then at the next most likely and so on).

We will need to define $\text{MIN}_{\preccurlyeq_i}(u)^n$ as follows:

$$\begin{aligned} \text{MIN}_{\preccurlyeq_i}(u)^0 &= \{\text{MIN}_{\preccurlyeq_i}(W_i^u)\} \\ \text{MIN}_{\preccurlyeq_i}(u)^{m+1} &= \text{MIN}_{\preccurlyeq_i}(u)^m \cup \{\text{MIN}_{\preccurlyeq_i}(W_i^u - \text{MIN}_{\preccurlyeq_i}(u)^m)\} \end{aligned}$$

That is, we first take the most plausible states; then we in effect 'break' the link to the most plausible states and look at the resulting states; and so on. As we continue along this process, it produces a set of nested sets (events); that is what Fact 3.7 states.

Fact 3.7. *For any $m, m' \in \mathbb{N}$, we have:*

1. *For all $X, Y \in \text{MIN}_{\preccurlyeq_i}(u)^m$, $X \subseteq Y$ or $Y \subseteq X$.*
2. *If $m' > m$ then $\text{MIN}_{\preccurlyeq_i}(u)^m \subseteq \text{MIN}_{\preccurlyeq_i}(u)^m$.*

Then each $\text{MIN}_{\preccurlyeq_i}(u)^m$ is rather like a set of 'belief spheres', to use terminology from Lewis [1973], which he introduced in the context of semantics for counterfactuals, that are very close to the plausibility ordering semantics we are considering for belief revision.

Furthermore, since we are considering finite models[5] this process will stop at some finite stage $m_i^u \in \mathbb{N}$, i.e. with $\text{MIN}_{\preccurlyeq_i}(u)^{m_i^u} = \text{MIN}_{\preccurlyeq_i}(u)^{m_i^u+1}$.

Now player i being lexicographically rational at u is going to mean i playing optimally with respect to the restrictions defined by *all* of the elements of $\text{MIN}_{\preccurlyeq_i}(u)^{m_i^u}$.

Definition 3.12. The event that i is ***lexicographically rational*** is the following one:

$$\mathbf{lr}_i := \{u \in W \mid \forall X \in \text{MIN}_{\preccurlyeq_i}(u)^{m_i^u},\ \xi_i(u) \in O_i(\xi(X))\}.$$

Prima facie this definition looks arbitrary, and for some optimality notions it might seem problematic. Let us first of all note that *for monotonic optimality operators* lexicographic rationality and rationality *tout court* (i.e. the notion of rationality that we have worked with up until now) are *the same*.

Fact 3.8. *If O_i is monotonic, then on any plausibility model,*

$$\mathbf{r}_i = \mathbf{lr}_i.$$

We should also look at a concrete example of an optimality operator in order to motivate this definition of lexicographic rationality. The only persistently non-monotonic

[5]The extension to the infinite case would be unproblematic: one would simply take the union at limit stages.

optimality operator that we have seen is *admissibility* (avoidance of weakly dominated strategies). Let us recall the definition of weak dominance (cf. Section 1.1). $wd_i(s_i, s_i', S_{-i})$ means that s_i is weakly dominated by s_i' with respect to S_{-i}:

$$\begin{array}{rll} wd_i(s_i, s_i', S_{-i}) & \text{iff} & \forall s_{-i} \in S_{-i}, (s_i', s_{-i}) \geq_i (s_i, s_{-i}) \\ & \text{and} & \exists s_{-i} \in S_{-i}, (s_i', s_{-i}) >_i (s_i, s_{-i}). \end{array}$$

Then the (contracting, global version of the) operator $nwd_i(S)$ is defined as

$$nwd_i(S) \;=\; \{s_i \in S_i \mid \forall s_i' \in T_i, \neg wd_i(s_i, s_i', S_{-i})\}.$$

Notice that our definition of lexicographic rationality is sensible in this context because of Fact 3.9.

Fact 3.9. *If* $wd_i(s_i, s_i', A_{-i})$ *and* $wd_i(s_i', s_i, B_{-i})$ *then* $A_{-i} \not\subseteq B_{-i}$.

Now, if Fact 3.9 did not hold, or if a players' plausibility ordering did not induce *nested* sets (Fact 3.7), then things would be problematic, because a player could then have two sets $A_{-i}, B_{-i} \in \text{MIN}_{\preccurlyeq_i}(u)^{m_i^u}$ where s_i is dominated by some strategy s_i' with respect to A_{-i}, with s_i' in turn being dominated by s_i with respect to B_{-i}. However, this situation cannot arise, and so our definition of lexicographic rationality is safe. Indeed, Fact 3.9 expresses an additional condition, on top of that given in Definition 1.14, that we might want to impose on optimality operators.

To get a little more concrete, let us see how lexicographic rationality can be used to define a sensible rational equilibrium of beliefs that *can* be reached even in the case of non-monotonic optimality operators. This time though we are not interested in *hard* public announcements of rationality or anything else, but in so-called '*soft*' announcements, of lexicographic rationality.

Where a hard public announcement models the passage of *hard* information, soft public announcements model the flow of *soft* information, and so is about changes of belief, where belief is modelled by a plausibility ordering as in the models we have just described.

The terminology of soft announcements is from [Benthem, 2007a], where a number of such operators are studied and axiomatised. One of those operators, that we will focus on, is called 'lexicographic update'. The idea of this epistemic action is that it is a soft announcement of some sentence φ, that has the effect that all those who hear it make *all* of the states where φ holds more plausible than those where φ does not hold, and *otherwise leave the ordering the same*. We refer to [Benthem, 2007a] for more discussion of this operation and justifications of it as a rational way of changing beliefs.

To close this Chapter, we will sketch how such soft announcements of lexicographic rationality can be used to generate a model that does explain, in our view better than the in the *hard* case, why players only choose strategies that survive iterated admissibility.

Just as in Definition 3.7, we want to say when a certain strategy *would* be rational for a player to play at a certain state. This time the definition of what would be rational is a little more involved, so we introduce some intermediary notation for this purpose.

Definition 3.13. The event that i's strategy $s_i \in T_i$ ***would be lexicographically rational*** for i, written $\mathbf{lr}_i(s_i)$, is as follows:

$$\mathbf{lr}_i(s_i) := \{u \in W \mid \forall X \in \text{MIN}_{\preccurlyeq_i}(u)^{m_i^u},\ s_i \in O_i(\xi(X))\}.$$

Now we can define an analogue to Definition 3.7, this time for plausibility models.

Definition 3.14. There is ***rational equilibrium at*** u ***in the plausibility model*** $\mathcal{M}$ just if:

$$\forall i \in N,\ \forall s_i \in T_i\Big(\mathbf{lr}_i(s_i) \subseteq \bigcap_{j \in N-\{i\}} [\![\Diamond_j s_i]\!]_{\mathcal{M}}\Big).$$

Definition 3.14 is susceptible to the same charges as Definition 3.7. Furthermore, the reader might prefer a strictly stronger condition, that takes into account not only players' unconditional beliefs, but also their conditional beliefs about other players' strategies. Again, we do not insist that the detail of Definition 3.14 is necessarily correct; the idea underlying it is the most important thing in order to make this largely conceptual point.

We do not discuss the syntactic aspects of lexicographic rationality, so let us give a purely semantic description of what is involved in arriving at a rational equilibrium of beliefs even in the non-monotonic case.

Start with the initial model $\mathfrak{I}'_G$ of some game G:

Definition 3.15. Given some game $G = (T, \geq)$, let $\mathfrak{I}'_G$ be the (pure) plausibility model $(T, \preccurlyeq_i = T \times T, id)$ in which the states are the strategy profiles of G, and all players have the same plausibility ordering $\preccurlyeq = W \times W$, so there are no non-trivial beliefs in this model.

Again: as in the case of $\mathfrak{I}_G$, the initial model $\mathfrak{I}'_G$ is intended to represent the epistemic situation of the players as soon as they have been presented with the game. So the game is common knowledge, but other than that the players have no information, soft or hard.

Now if we iteratively *softly* announce *lexicographic* rationality, then we will end up with a model very much like that depicted in Figure 3.5 above, except that the 'partitions' there should be nested.[6] It is easiest to draw the plausibility ordering $\preccurlyeq_i$ at u in terms of the induced 'sphere system' $\text{MIN}_{\preccurlyeq_i}(u)$, and that is what we do in Figure 3.9 in order to represent the model that is generated by soft announcements of lexicographic rationality. In this way we generate a model in which the *only* rational strategies that can be chosen are those that survive the iterated elimination of weakly

[6]Iterations of different kinds of soft announcements are studied, and some interesting properties concerning cycles and fixpoints are found, in [Baltag and Smets, 2009].

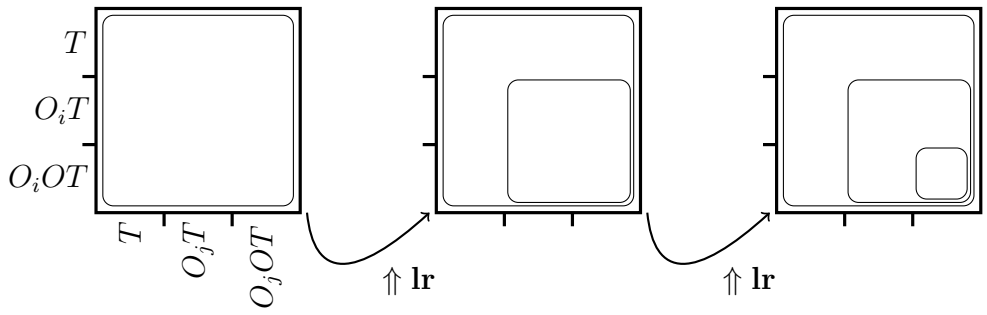


Figure 3.9: Soft announcements of lexicographic rationality. The picture here is like Figure 3.5, except that a belief ordering, rather than a partition, of the state space is being formed.

dominated strategies. The model really does represent a *rational equilibrium of beliefs*: no player will be tempted, as we saw would be the case with the previous 'hard information' analysis, to not play according to the predictions of the equilibrium, since there is always a reason for acting in accordance with it.

We have not stated any significant results in this Section, which has primarily consisted of definitions and remarks about them. One aim of these definitions has been to motivate the use of plausibility models, and so conditional beliefs, in the analysis of deductive reasoning in games, in order to understand the dynamics of non-monotonic optimality operators on strategic games. We will actually put some of these definitions to more substantial use in the next Chapter when we apply a similar kind of logical analysis to extensive games.

Summary

In this Chapter we have looked at the dynamics of information, including some applications to themes from the previous two Chapters. We presented dynamic epistemic logic (DEL), along with some minor new results concerning neighbourhood models. We used DEL to analyse game reduction processes themselves, rather than just looking at the results of those processes. We tried to tell a coherent story about that process, and showed also how models like those from the Theorems from Chapter 1 could be built by repeating some action, corresponding to the process of deliberation that players go through in the one-shot interaction scenarios we are considering.

This is a formalisation of the private but common process that underlies deductive reasoning about games. We also introduced the notion of a 'rational equilibrium of beliefs': the idea is that although we are not in what game-theorists might call equilibrium (according to some steady-state interpretation), there still should be an equilibrium of a kind, and this led us to suggest the significance of conditional beliefs in understanding non-monotonic solution concepts.

Introducing conditional beliefs also meant that we have already laid some of the technical foundations for the next chapter, where we will make extensive use of the plausibility models defined here.

Chapter 4

Extensive Games

"*Aumann has proved that common knowledge of substantive rationality implies the backward induction solution. Stalnaker has proved that it does not.*"
– Joseph Y. Halpern [Halpern, 2001]

The only games we considered in the preceding chapters were strategic games. What about applying our logical analysis to extensive games?

Extensive games represent decision processes that are extended over time, and so players do not always make their choices in complete ignorance about what choices other players make.

We start by looking at games of so-called 'perfect information' (PI), in which every move by every player is commonly and synchronously observed. In such games backward induction, the "oldest idea in game theory" [Aumann, 1995, p. 635], yields a natural solution concept, subgame-perfect equilibrium. There has been a great deal of debate concerning just what would be a sound epistemic foundation for it. One of the aims of this chapter is to offer a perspective on that debate that is provided by epistemic or doxastic logic. We use a logical language for conditional belief in order to express some of the key notions that we will use, including a *dynamic, forward-looking* form of rationality. We model the *actual playing of the game* within our logical language using *public announcement modalities*: when a node p in the tree is reached, the intuitive description of a PI game means that we can legitimately say this is a common collective learning *that* p has been reached. We then prove Theorem 4.1 that gives epistemic conditions, in terms of dynamic rationality and common conditional belief, for backward induction.

Backward induction in so-called 'generic' PI games has the pleasant feature that it yields a unique outcome. So *purely* by reasoning about the game, from the Deductive interpretation that we have in mind in this Thesis, players can arrive at a unique

outcome. Subgame-perfect equilibrium is a refinement of Nash equilibrium, which depends upon a steady-state interpretation. Nonetheless in generic games the backward induction algorithm allows that the unique subgame-perfect equilibrium be derivable by each player in isolation (as we might say, repeating a phrase from Chapter 3: 'privately but commonly') reasoning about the game situation.

We try to take an approach to the questions that underlie the problem of backward induction that is as generic as possible, and as untainted by ambiguously-interpretable game-theoretical notions. So in particular, we do not (explicitly) make extensive-form *strategies* into the objects of players' beliefs. Indeed, in the formal language that we introduce for reasoning about models for games in extensive form, we will not include terms for strategies, but for *outcomes* or (equivalently) for *nodes* in the game tree.

We look briefly at games with imperfect information, and illustrate how DEL action models (introduced in Section 3.1) could be used to simulate the play of an extensive game of imperfect information (where PI games were simulatable just with public announcements). We will also look at a *limitation* of that DEL-based analysis: that is not, as it stands, fit to analyse the interesting phenomenon of strategic communication.

Selten, who introduced the notion of subgame-perfect equilibrium, later [1975] introduced a further refinement of it, which he called 'perfect equilibrium'[1], and which is often now called 'trembling-hand equilibrium' (e.g. in [Osborne and Rubinstein, 1994]). Epistemically speaking, this concept requires a great deal from the steady-state interpretation of game theory. Not only must players be correct about each other's strategies, but what is more they must in effect share a conditional probability system specifying with what probabilities each event pertains should various other events (including those they actually believe to hold) not in fact hold. We offer a simplified version of trembling-hand equilibrium, 'even-handed trembling-hand equilibrium', that supposes only that players are correct about each other's strategies, and share a very natural belief revision policy.

Background literature

The issues that we deal with in this Chapter originate in the work of Aumann [1995], Stalnaker [1994; 1996; 1998] and Reny [1992], and have been investigated by a number of authors: [Binmore, 1987; 1996; Bicchieri, 1989; Battigalli, 1997; Battigalli and Siniscalchi, 1999; 2002; Bonanno, 1991; Brandenburger, 2007; Halpern, 2001; Samet, 1996; Clausing, 2003] is not an exhaustive list, and there are many illuminating discussions to be found in the literature. Many of the different solutions proposed in those works are related in different ways to our own.

The arbitrary announcement modality $[!]\varphi$ is introduced and studied in [Balbiani *et al.*, 2008].

[1] He has previously used the term 'perfect equilibrium' for what is now known as 'subgame-perfect equilibrium', and so remarks that "In retrospect the earlier use of the word 'perfect' was premature" (op.cit.).

Organisation of the Chapter

In Section 4.1, we introduce our notation for extensive games, and define many of the important concepts like strategies, backward induction, conditional belief models for extensive games, etc.

In Section 4.2, we will go on to present conditions that ensure that players will play according to the backward induction outcome.

Finally, in Section 4.3, we look at extensive games with imperfect information. We also look at Selten's 'perfect equilibrium' or 'trembling-hand equilibrium', which was introduced for extensive games, but is most commonly defined for strategic-form games (for example in [Osborne and Rubinstein, 1994]). We therefore move back to a strategic game perspective to develop 'even-handed trembling-hand equilibrium'. We close the Chapter by raising the question how to analyse strategic communication using our formal epistemological methodology.

4.1 Games with perfect information

Extensive games differ from the 'strategic' or 'normal-form' games that we have so far considered in that they are supposed to represent an extended process of play, in which players take it in turns to make moves, rather than choosing a single strategy and sitting back and waiting. Let us consider why it is normal that we should add something to the ideas from the previous chapters before using them to reason about games in extensive form, given that these represent a process that is extended in time. It is possible to define *strategies* for extensive games, that can be used to define a strategic game that is in some sense equivalent to the extensive game. That is, one could simply translate an extensive game into its normal form (see the definition in Section 4.1 below), and then apply an existing analysis, in terms of public announcements (Chapter 3) or in terms of common belief of rationality (Chapter 1), to the resulting game. However, this will not be very revealing in the sense that it will not yield any insight in or understanding of the subtleties of the information dynamics of extensive games.

As a number of commentators have observed (for example [Stalnaker, 1996; Bruin, 2004]), treating extensive games as if they were strategic games in this way means ignoring the crucial feature of extensive games. After all, an extensive game is supposed to represent a multi-party *sequence* of decision processes extended over a temporal interval, i.e. one player makes his move after another player has made hers. In order to do justice to this natural interpretation, we need to allow for players' beliefs to change *as the game is played out*. That is, we have a "many-moment interpretation", and not a "one-shot interpretation" of strategic games in mind (cf. [Bruin, 2004, Chapter 4]).

Furthermore, it is not just that players should be able to (monotonically) *increase* their beliefs, but actually to *revise* them. If for example, in a game of chess, player 1 believes that player 2 will advance his queen's pawn but he instead castles, we certainly do not want player 1 to maintain her previous belief concerning player 2's move. For

players to be able coherently to revise their beliefs as the game goes on, we will in this Chapter again use conditional belief models.

Even if we do have a "many-moment" interpretation in mind for the particularities of extensive games, still this is independent of the Deductive interpretation of games, which we want to maintain in analysing extensive games of perfect information. So the main question is (still) what the players can deduce from some basic common principle of rationality.

An extensive game is based around a (finite) game *tree*, which consists of a relation $\twoheadrightarrow$ over a finite set $\mathcal{Z}$ of *nodes*;

Definition 4.1. For it to be a tree, $\twoheadrightarrow$ has to be reflexive transitive and *anti*symmetric[2], and for each $p \in \mathcal{Z}$, restricted to the set $\{q \mid q \twoheadrightarrow p\}$, the relation $\twoheadrightarrow$ is *total*: i.e. for each $q, q' \in \mathcal{Z}$, if $q \twoheadrightarrow p$ and $q' \twoheadrightarrow p$, then either $q \twoheadrightarrow q'$ or $q' \twoheadrightarrow q$.

For $p, q \in \mathcal{Z}$, $p \twoheadrightarrow q$ means that q can be reached by descending[3] branches of the tree from p. we write $p \mapsto q'$ to mean that q' is an *immediate* successor of p (i.e. that $p \twoheadrightarrow q'$ and for no other $q \neq p$ and $q \neq q'$ is it the case that $p \twoheadrightarrow q \twoheadrightarrow q'$.

We in general also lift these relations to functions, though we use the convention that $\hookleftarrow (p)$ to denote the unique node q such that $q \mapsto p$ (rather than the singleton set containing it, as the usual lifting of relations to functions would have it).

In a tree there is a single 'root' node, one that has no predecessor. We generally denote this node by r. Each non-terminal node of the tree is assigned to a particular player, via a function $\rho : \mathcal{Z} \to N$. The player $\rho(p)$ is the player who's turn it is at p. We think of her as being in the situation of having to choose some $q \hookleftarrow p$.

The terminal nodes, or 'leaves', of a game tree are called *outcomes*: once every player has made her choice, at every node that is reached, then we are in the same situation as in the case of a strategic game when all players have chosen their strategy. Thus in a game each player will have preferences over these outcomes.

Formally we define the leaves of a tree as follows:

Definition 4.2. The ***leaves*** (or ***outcomes***) of a tree $(\mathcal{Z}, \twoheadrightarrow)$, are written $\mathcal{O}(\twoheadrightarrow)$, where:

$$\mathcal{O}(\twoheadrightarrow) = \{p \in \mathcal{Z} \mid \forall q \in \mathcal{Z}, \neg(p \twoheadrightarrow q)\}.$$

It will be convenient at times to talk about the 'edges' of the tree, i.e. the *actions* that players take at different nodes in order to move between them. Formally speaking, these are the pairs $(p, q) \in \mapsto$. We use a *labelling* function to label the actions with names. Given some set L of labels, a labelling function ℓ associates with each node, other than the root, a label. The idea is then that $\ell(p)$ is the name for the action that just occurred at p. We can in the standard way lift this function to a set, and so we can write $\ell(\mapsto (p))$ to mean the set of labels of actions available at the node p.

[2]We have already seen those first two properties of relations; the last means that if $p \twoheadrightarrow q$ then either $p = q$ or $\neg(q \twoheadrightarrow p)$.

[3]Game-theorists usually draw trees going downwards.

In some definitions of extensive games, edges are taken to be primitive, and points (or 'histories') are defined in terms of them. We prefer to take points as primitive (and indeed later in Section 4.2 will think of trees as sets of sets of outcomes).

In this Section and the next, we are interested in extensive games of so-called 'perfect information'. What that qualifier means is that when a node p in the game tree is reached, all players correctly believe that p has been reached. In fact, here when we come to the epistemic analysis of games, we will allow ourselves to say that players 'know' where they are in the tree, since the actual play of the game, as it occurs, really is taken to be an observable; recall the quotation from the Introduction, that we endorse: "only observables are knowable [...] and only moves are observables" [Brandenburger, 2007].

There is no need to add any informational structure to the tree itself then, until we later define extensive games with imperfect information.

Definition 4.3. ***Extensive games of perfect information*** (sometimes '***PI games***', or even just 'games') for the players N are those structures of the form

$$(\mathcal{Z}, \twoheadrightarrow, \rho, \leq_i)_{i \in N},$$

where for each $i \in N$, $\leq_i$ is a total linear order over the outcomes $\mathcal{O}(\twoheadrightarrow)$. If $\leq_i$ is *strict*, i.e. if each player has a preference between every outcome, then we call the game ***generic***.

Another piece of notation will be useful: given some player i, we denote by $\rho_i(\mathcal{Z})$ the set of nodes in $\mathcal{Z}$ where i plays, i.e.:

$$\rho_i(\mathcal{Z}) = \{p \in \mathcal{Z} \mid \rho(\mathcal{Z}) = i\}.$$

Note that we do not explicitly include the labelling function ℓ in the definition of a game of perfect information. Still it will be useful for us to talk about the actions (edges), and so sometimes we assume that there is some labelling function associated with a given game. This choice is made simply because we do not actually need to include a labelling function. (When, in Section 4.3, we consider games of imperfect information, we will include a labelling function in the definition.)

A *strategy* for player i in a game of perfect information has to tell i what to do no matter what happens.

Definition 4.4. A ***strategy*** s_i is a function from $\rho_i(\mathcal{Z})$ to $\mathcal{Z}$ such that $\rho_i(p) \hookleftarrow p$.

A strategy profile is then, as in the case of strategic games, a profile $(s_i \mid i \in N)$ of strategies, one for each player. Notice that a strategy profile s is fully deterministic, in the sense that it will determine a unique outcome, that we write $\mathfrak{o}(s)$: $\mathfrak{o}(s)$ is the unique element of $\mathcal{O}$ such that there is some $k \in \mathbb{N}$ with $s^k(r) = \mathfrak{o}(s)$ (where r is the root of the game tree).

In fact though, strategies are *more* than just fully deterministic plans: they go beyond what is actually needed by a player. For instance, consider the one-player game

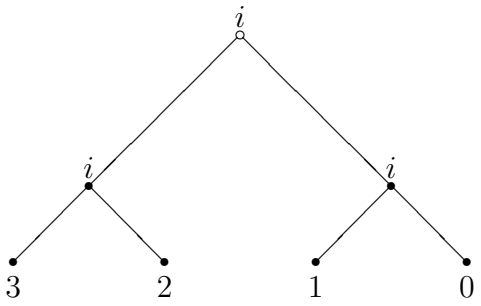


Figure 4.1: A game illustrating that strategies are more than plans

illustrated in Figure 4.1 (where, as in the case of strategic games, we represent the ordinal preferences in the game with numbers). The game situation could represent the following situation. Assume, for the sake of argument, that the player, i, cares a great deal about money, and so prefers, other things being equal, a situation where he has more money to a situation where he has less money. Then suppose the situation is the following: player i is going to be offered two units of money, and then one unit of money, and can either accept (play the left-hand node), or refuse (play the right-hand node). Accepting or declining in the first round has no effect on the game other than that the player will receive less money. Then this situation is represented by the game in Figure 4.1: if player i accepts twice, that is the best situation, followed by the situation in which he accepts first and then declines, following by the situation in which he declines first then accepts, and the worst situation (for i) is that where he declines the money both times, for then he has got no money out of the interaction.

Now, the obvious way for player i to play this game (which is really just a decision problem, since there are no other players in it) is to accept both times. So we might like to say that his strategy s_i *is* to play 'accept' and then 'accept' again. But having a complete strategy, according to the definition above (standard from game theory) means also having a sort of 'counterfactual' plan: i's strategy must also assign a successor to the node arrived at by playing right, i.e. the node that would be arrived at if i were to decline the money, *even if i's strategy says he should accept the two units of money offered in the first round*.

Now that we have defined strategies in extensive games, in such a way that a strategy profile uniquely determines an outcome, there is an obvious translation from extensive games to strategic ('normal form') games. Given an extensive game, the strategic game induced by it has as (normal-form) strategies the (extensive-form) strategies from Γ, and the preferences in the strategic game are just the preferences over the outcomes determined by the given (extensive-form) strategy profile.

Definition 4.5. The ***normal form of an extensive game*** $\Gamma = (\mathcal{Z}, \twoheadrightarrow, \rho, \leq_i)_{i \in N}$ is $(T_i, \leq_i')_{i \in N}$, where

$$T_i = \{s_i \in \mathcal{Z}^{\rho_i(\mathcal{Z})} \mid s_i(p) \hookleftarrow \mathcal{Z}\},$$

and

$$s \leq' s' \text{ iff } \mathfrak{o}(s) \leq_i \mathfrak{o}(s').$$

The next solution concept that we will introduce is an explicitly extensive-form notion. However, note that it is also known, in the case of generic perfect information games that we will consider, to be equivalent to iterated admissibility. The epistemic condition for backward induction that we propose in Section 4.2 is also phrased in terms of *conditional beliefs*. We suggested that the logic of conditional beliefs is what best explicates the stability of the 'rational equilibrium of beliefs' achieved with respect to iterated admissibility, and there is some similarity between what we suggest here and that approach. Still, there will still be a uniquely extensive-form flavour to the condition.

Before defining subgame-perfect equilibrium, we need to define the notion of a subgame: The term 'subgame' is used, for extensive-form games, with a different meaning from that intended in the case of normal-form games. If p is a node in the perfect-information game Γ, then the (extensive-form) *subgame generated by* p is just the *restriction* of Γ to the nodes reachable by descending in the tree from p (so including p itself).

Definition 4.6. If

$$\Gamma = (\mathcal{Z}, \twoheadrightarrow, \rho, \leq_i)_{i \in N},$$

then the subgame generated by p is the following game:

$$\Gamma^p = (\twoheadrightarrow(p), \twoheadrightarrow \restriction \twoheadrightarrow(p), \rho \restriction \twoheadrightarrow, \leq_i \restriction \twoheadrightarrow)_{i \in N},$$

where $F \restriction X$, read 'the restriction of F to X' is the unique function/relation with domain restricted to X that agrees, everywhere on X, with F.

Subgame-perfect equilibrium is a refinement of *Nash equilibrium*, that we talked about informally in the Introduction. Let us define Nash equilibrium properly now.

Definition 4.7. A strategy profile s in the N-player strategic game $(T_i, \geq_i)_{i \in N}$ is a ***Nash equilibrium*** just if for all players i, s_i is a best response, among T_i, to s_{-i}. That is:

$$\forall i \in N \, \forall s_i' \in T_i, \; s \geq_i (s_i', s_{-i}).$$

Recall that this is indeed an 'equilibrium' notion, so accords with the steady-state interpretation of games, in the sense that the epistemic justification for playing according to s is that the players believe that the others will play according to s, and therefore have no reason to switch and play another strategy $s_i' \neq s_i$. Of course, a player i might be indifferent between two such strategies, in which case she might not play according to a given Nash equilibrium, but it is always *rational* for her to play according to the equilibrium.

Proposition 4.1 (cf. [Aumann and Brandenburger, 1995, Preliminary Observation])**.** *If all the players are rational and have a correct belief about the other players' strategy choices, then they play a Nash equilibrium.*

Given some language that can define Nash equilibrium (cf. [Benthem *et al.*, 2006]) as some sentence NASH, as well as rationality and belief, Proposition 4.1 is equivalent to the validity of the following sentence:

$$\mathbf{s} \rightarrow ((\Box \mathbf{s} \wedge \mathbf{r}) \rightarrow \text{NASH}).$$

The idea of a *subgame-perfect* equilibrium is just that is a Nash equilibrium in *every subgame* of the game:

Definition 4.8. Given some extensive game $\Gamma = (\mathcal{Z}, \twoheadrightarrow, \rho, \leq_i)$, the strategy profile s is a ***subgame-perfect*** equilibrium iff in *each* subgame Γ^p generated by *every* node p of Γ, the strategy profile $s{\restriction}(\twoheadrightarrow(p))$ of that subgame is a Nash equilibrium.

There is always at least one subgame-perfect equilibrium, and in the case of generic games, there is a unique strategy profile that is a subgame-perfect equilibrium. Clearly any subgame-perfect equilibrium will be a Nash equilibrium, since the whole game is a subgame of itself (generated by its root node). To see that it is a strict '*refinement*' of the notion of Nash equilibrium, and to understand the intuitive justification of the definition, we give in Figure 4.2 an example of a game with a Nash equilibrium that is *not* subgame-perfect. We indicate a particular strategy profile in a picture of a game

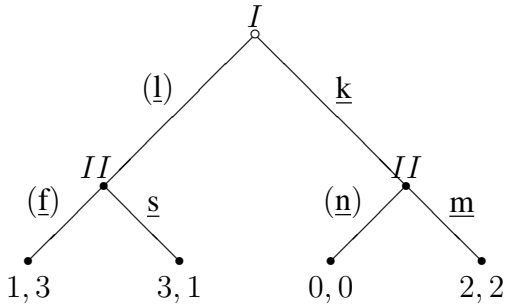


Figure 4.2: A game with a Nash equilibrium that is not subgame-perfect

by writing the relevant name $\ell \in L$ of the choices the strategy says to use (one for each non-terminal node) as (ℓ). The strategy profile indicated in Figure 4.2, in which I plays left, and II in fact plays left *and according to her strategy* would *play left* if II were to play right, is a Nash equilibrium. Deviating for player II means either choosing the right-hand node in the game as it is actually played, and thereby getting a less preferred outcome ($1 < 3$), or switching the hypothetical play to play right. This *would* guarantee that II obtains 2 rather than 0, *if* I were also to deviate, but we are keeping I's strategy fixed (as the definition of Nash equilibrium requires), and so this

other deviation of II has no effect. Notice also that player I has an incentive *not* to deviate, since the only way to deviate for I would be to choose the right instead of the left node. But I plays the right node, then, since II plays the left node either way (according to the strategy profile being considered), I would then get a less preferred outcome ($0 < 1$).

It is perhaps easier to verify that in the normal form of the game in Figure 4.2, that we give in Figure 4.3, the relevant outcome is a Nash equilibrium.

	nf	ns	mf	ms
k	$0, 0$	$0, 0$	$2, 2$	$2, 2$
l	$1, 3$	$3, 1$	$1, 3$	$3, 1$

Figure 4.3: The normal form of the game from Figure 4.2

The motivation behind subgame-perfect equilibrium should also be clear from this example as well. Although II's choice in the right-hand node does not affect the actual outcome, still it plays a role in the situation: it can be thought of as a 'threat', that player II *would* play to the $0, 0$ outcome if given the chance. Of course, when seen as such, it is not a *credible* threat: if actually confronted with the choice between $0, 0$ and $2, 2$, player II would be going against her own preferences if she were to choose 0. To put it in other terms: in the relevant subgame (in which I would have played right), II would not be playing a best response by sticking to the strategy that tells her to play left.

In the next Section, we will work towards establishing conditions on beliefs and rationality that entail that the players play a subgame-perfect equilibrium.

In generic games, like that in Figure 4.2, subgame-perfect equilibrium always yields a unique strategy. For the rest of this Section, and indeed until Section 4.3, unless we specify otherwise we assume the game to be generic. Now, that unique strategy can be computed via a simple algorithm, a process known as 'backward induction', so-called because it works 'backwards', going 'up' the tree (i.e. from the leaves to the root).

The algorithm works by tagging edges in the game tree, ending up with a unique edge (p, q) tagged for each $p \in \mathcal{Z}$. This tagging then yields the subgame-perfect equilibrium strategy.

The tag is built up step-by-step; we initially set the tag (the set of tagged edges) as $\emptyset$. Take a game $(\mathcal{Z}, \twoheadrightarrow, \rho, \leq_i)_{i \in N}$. Start at the end of the tree, at the leaves, and look, for each leaf p, at its predecessor, the unique q' such that $p \mapsto q'$. Then pick the unique leaf q, for each q', that is maximal with respect to $\leq_{\rho(q')}$, and tag (p, q), i.e. add (p, q) to the tag set. (This *unique* leaf is guaranteed to exist because the game is generic.) Why do we pick the maximal leaf for player $\rho(q')$? Because, as in Chapter 1, the hypothesis we are working with is that the players are rational, and a rational player is one who chooses her most preferred option. And since it is by definition $\rho(q')$ who plays at q', then we assume that the $\leq_{\rho(q')}$-maximal would be reached if q' were reached.

More formally:

$$\mathrm{BI}_0 = \{(\leftarrowtail (q), q) \in \mathcal{O}(\twoheadrightarrow) \mid \forall p \leftarrowtail\!\!\mapsto q,\ p \leq_{\rho(\leftarrowtail(q))} q\}.$$

That is the initial step of the 'inductive' process. Part of our inductive hypothesis (which clearly holds at the BI_0 level, and can be verified to hold after this successor inductive step) will be that for every $(q, q') \in \mathrm{BI}$, there is a unique path $q = q_0 \mathrm{BI}_m q_1 \mathrm{BI}_m \ldots \mathrm{BI}_m q_m \in \mathcal{O}(\twoheadrightarrow)$. We write $\overrightarrow{\mathrm{BI}}(q)$ to mean the unique $q_m \in \mathcal{O}(\twoheadrightarrow)$ that is reachable by a tagged path from q. For the inductive step then, we now look, instead of at the leaves, at those nodes p such that

1. there is no tagged edge $(p, q) \in \mathrm{BI}_k$, but
2. for every $q \leftarrowtail p$, there is an edge $(q', q) \in \mathrm{BI}_k$.

We write X_k for the set of non-terminal nodes $p \in \mathcal{Z} - \mathcal{O}(\twoheadrightarrow)$ satisfying these conditions 1. and 2. If X_k is not empty, then we proceed to define BI_{k+1}. Since $X_k \neq \emptyset$, there are some p's satisfying 1. and 2. above. On each of these we perform a similar algorithmic procedure to that carried out on the leaves in order to define BI_0, that is, for each node p, we find the unique successor $q \leftarrowtail p$ of it that leads, by following the BI_k relation to the $\geq_{(} \rho(p))$-maximal outcome among all outcomes that are reachable by going to any successor $q' \leftarrowtail p$ and then following the BI_k relation. That is, for each p, if i is the player choosing at p then we pick (on i's behalf, so to speak) the unique q such that for any $q' \leftarrowtail p$, $\overrightarrow{\mathrm{BI}_k}(q) \geq_i \overrightarrow{\mathrm{BI}_k}(q')$. The corresponding edge for each such node that we pick is then added to BI_k, forming BI_{k+1}:

$$\mathrm{BI}_{k+1} = \mathrm{BI}_k \cup \{(p, q) \in X_k \times \mathcal{Z} \mid p \mapsto q \text{ and } \forall q' \leftarrowtail p,\ q' \leq_{\rho(p)} q\}.$$

Because we are working backwards through the tree, if $X_k = \emptyset$, then it is because condition 1. fails, in which case if we have performed in total k of these successor steps, then we have defined BI_k, which is the subgame-perfect strategy profile:

Fact 4.1. *If $X_k = \emptyset$ then for every $p \in \mathcal{Z} - \mathcal{O}(\twoheadrightarrow)$, there is some $q \leftarrowtail p$ such that $(p, q) \in \mathrm{BI}_k$. I.e. BI_k is a strategy profile. Furthermore, by construction BI_k is subgame-perfect.*

Since the game is finite, this process will terminate at some stage k (i.e. for some $k \in \mathbb{N}$, $X_k = \emptyset$). We write BI for the resulting tagged set of edges, which is the subgame-perfect equilibrium strategy profile. $\mathfrak{o}(\mathrm{BI})$, the outcome reached by following this strategy profile, is called the 'backward induction outcome'.

This algorithm, like the iterated elimination of non-optimal strategies in a normal-form game, can also be thought of as being some 'private but common' process: working just from a basic notion of rationality, all players can perform the algorithm, and are legitimated in doing so on the basis that the other players are performing it.

The backward induction algorithm has been analysed in terms of public announcements in [Benthem, 2007b], where it is shown that, again, repeated announcements

of rationality, suitably defined for a model of the game tree, will 'prune' the model (the tree), and yield the backward induction outcome. That work is an interesting tangent to the main currents of debate in the epistemic game theory literature, which have centred around a controversy as to what the correct epistemic conditions for backward induction (subgame-perfect equilibrium) really are.

In generic games, the backward induction *outcome* is uniquely determined by iterated admissibility. That is: take the normal form of an extensive game, and iteratively eliminate all weakly dominated strategies. Then a region of the normal-form game will be left, that does not necessarily determine a unique strategy profile, but all of the strategy profiles that are left will determine the same outcome, and that outcome will be the backward induction outcome (cf. [Duggan, 2003]).

So one way to give the epistemic conditions that explain the backward induction *outcome* would be to give conditions that explain iterated admissibility. However, as we have seen that is not entirely straightforward, and furthermore does not in our view get to the heart of the matter. Recall that when until we got the analysis right, the reasoning behind iterated admissibility undermined itself (cf. Section 3.3). As we will see in the next Section, there is a similar sort of non-monotonicity in the reasoning behind backward induction: the reasoning appears, on first view, to undermine itself. Therefore, just as we argued that *conditional beliefs* are important in explaining why players play according to iterated admissibility, so we will use conditional beliefs in analysing backward induction.

In the next Section, we give conditions in terms of conditional beliefs, and a concept of rationality that is specific to extensive-form games, that guarantee that players play according to the backward induction outcome. Furthermore, when the players' *conditional beliefs* are interpreted as being about their strategies, then we can read the conditions we give as entailing that the *strategy profile* chosen is the subgame-perfect equilibrium profile.

4.2 Conditions for backward induction

In this Section we enter the debate about what the epistemic conditions for backward induction really are. First we mention some early results about the epistemic foundation for backward induction, and see the paradox involved in the reasoning behind backward induction. Then we will present conditional logic models for extensive games introduce a formal language for reasoning about them, and see how public announcements can be used to model moves in games (of perfect information). We will then formulate epistemic conditions that are sufficient for backward induction, and that we believe are truly an explanation of the issues involved in backward induction. One advantage of our approach is that we do not include explicitly in our formalism the conceptually problematic notion of *strategy*, with its counterfactual connotations. Instead, strategies will be a derived notion.

Our main contribution in this Section is to give conditions that are sufficient for

backward induction, and that we claim do justice to the conceptual issues involved.

Aumann [Aumann, 1995] has proved within the context of a partition-space model that common *knowledge* of 'rationality', entails the backward induction outcome, and that such a model will always exist. There rationality means something very strong, what is known as 'substantive rationality', something like *choosing optimally everywhere* in the tree. Needless to say, there is nothing wrong with his formal argument; but we would like to suggest, along with [Stalnaker, 1996; Samet, 1996] among others, that conceptually speaking his whole framework does not do justice to the problematic issues involving counterfactuals, or indeed take into account possible *changes* or *revisions* of belief during the game. There is a substantial literature in which the view we take of the deficiencies of the knowledge-based analysis is expressed [Reny, 1992; Binmore, 1987; 1996; Bonanno, 1991; Bicchieri, 1989; Brandenburger, 2007].

The reasoning that underlies the backward induction method seems to give rise to a paradox: in order even to start the reasoning, a player assumes that true common belief in "rationality" holds. So in particular (so the player is supposed to reason) at all of the last decision nodes $\leftarrow\!\!\!\shortmid (\mathcal{O})$, i.e. those just before the leaves, there is rationality. This entails that the obviously irrational *leaves* are eliminated. However, in the next reasoning step (going backward along the tree), some of these (last) decision nodes, some subset $Y \subseteq \leftarrow\!\!\!\shortmid (\mathcal{O})$ will be eliminated, on the same basis: that they are incompatible with (common true belief in) "rationality". But then the assumption behind the previous reasoning step is now undermined: the reasoning player can now see, that *if* those decision nodes Y that are now declared "irrational" were ever to be reached, then the only way that this could happen is if (common true belief in) "rationality" failed. Hence, the player was *wrong* to assume (common belief in) "rationality" when she was reasoning about the choices made at those last decision nodes. So, in a manner reminiscent of iterated admissibility, the whole line of argument seems to undermine itself.

Consider as an example the "centipede" game (cf. [Rosenthal, 1981]) given in Figure 4.4. This is a two-player game for a (Alice) and b (Bob). The reason this is called

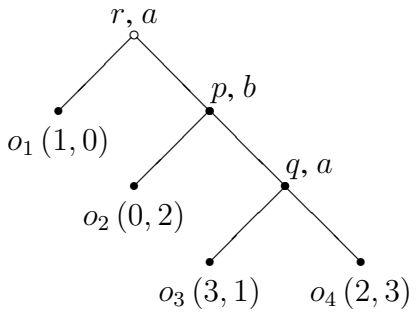


Figure 4.4: A "centipede" game

the "centipede" is that it can be extended indefinitely (so also one hundred times), to

arrive at a game where the backward induction outcome yields a very low payoff for both players as they 'exit' the game (a plays to o_0 and indeed at all nodes the unique subgame-perfect equilibrium says that each player will always take the left-hand option), compared to the payoff they could achieve by staying in the game. The small game in Figure 4.4 will be our running example in this Section, so let us take a few moments to become familiar with this example, that is relatively familiar from the literature. Here the backward induction algorithm does the following: at the end player a is choosing, and so chooses the most preferred option, i.e. o_3. Therefore at p, player b is choosing (to attribute 'choice' and player b's agency to the algorithm) between o_2 and $\overrightarrow{\mathrm{BI}_0}(q)$, i.e. between o_3 and o_2 and o_3, and so picks o_2; similarly, when choosing between $\overrightarrow{\mathrm{BI}_1}(p)$, i.e. o_2 and o_1, at r, where $\rho(r) = a$, the algorithm picks o_1.

We will use 'impure' plausibility models (Definitions 3.9 and 3.11), with an operator for knowledge. However one thing we will *not* assume as known is the *future of the game*: no outcomes that are consistent with the structure of the game are to be excluded at the outset of the game. In fact, we make the opposite assumption: that it is common knowledge that nobody knows the future, i.e. nobody knows that some outcome will not be reached. This "open future" assumption seems to contradict common knowledge of rationality; but in fact, it is consistent with it, if by rationality we only mean "rational planning", leaving open the possibility that players may make *mistakes* or may change their minds. The players may certainly *believe* their rational plans will be faithfully carried out, but they have no way to *know* this in advance. We think of our "open future" assumption as being a realistic one, and moreover one that embodies the players' "freedom of choice", as well as the "possibility of error", that underlie a correct notion of rationality.

A player's rationality can be assessed only if she is given some options to freely choose from. There are certainly cases in which the future can be known, e.g. when it is determined by a known natural law. But it is an essential feature of rational players that their own choices are not known to them to be thus determined: otherwise they would have no real choices, and thus no rational *choice*. Any natural determinism is assumed to be absorbed in the definition of the game structure, which does pose absolute limits to choices. In a sense, this simply makes precise the meaning of our knowledge as that which is produced by hard information, and makes a strict delimitation between the past and the future choices, delimitation that is necessary in order to avoid the various paradoxes and vicious circles that plague the notions of rational decision and freedom of choice: the players may have hard information about the past and the present, but not about their own future free choices (although they may have "soft" information, i.e. "certain" beliefs, with probability 1, about their future choices).

Maybe the most important original feature of our analysis is the notion of *"dynamic" rationality* that we introduce, which takes into account the dynamics of beliefs, as well as the dynamics of knowledge. On the one hand, following Stalnaker, Reny, Battigalli and Siniscalchi and others (and in contrast with Aumann), we assess the rationality of a player's move at a node against the beliefs held *at the moment when the node is reached*.

On the other hand, we incorporate the above-mentioned epistemic limitation to rationality: the rationality of a player's move only makes sense when that move is *not already known* (in an irrevocable manner) to her. *Players cannot be held responsible for moves that they cannot choose or change any more (including their own past moves)*. Since the players' knowledge increases during a game of perfect information, their set of available options decreases: passed nodes, or nodes that were by-passed, cannot be the objects of choice any more. As a result, our notion of rationality is 'future-oriented': at any stage of the game, whether or not a player is dynamically rational at that stage depends only on her current and future moves.

So a player can be rational *now* even if in the past she has made some "irrational" moves. In effect, performing such an irrational move in a game of perfect information is in part a public announcement that "the player is (currently) *not* rational" (at the moment of moving). All the players jointly learn this fact (as a piece of hard information), but the fact itself may no longer be true after being learnt: while previously irrational (since about to make a 'wrong' move), the player may become rational after the wrong move (simply because, for all the decisions that she can still make after that, she chooses the 'right' moves). So the truth-value of the sentence "player i is (dynamically) rational" may change after a move by player i.

The way this is captured and explained in our formal setting is original and interesting in itself: the meaning of our "rationality" changes in time, due to the change of beliefs and of the known set of options. This is because the rationality of a player is an epistemic-doxastic concept, so it is affected by any changes in the information possessed by that player (including the changes induced by the player's own moves). In our setting, this is of course a natural and perfectly standard feature, an immediate consequence of the epistemic definition of rationality: epistemic sentences do not necessarily preserve their truth value after they are announced. An instance of this phenomenon is the 'Moore sentence' $p \wedge \neg \Box_i p$, which is *never* true after it is "learnt".[4]

Our concept of dynamic rationality, developed on purely a priori grounds, is at the heart of our resolution of the paradox of backward induction. Recall that the first reasoning step in the argument (dealing with the last decision nodes of the game) is no longer undermined by the result of the second reasoning step, since the notion of "rationality" assumed in the first step is not the same as the "rationality" disproved by the second step. The second step only shows that some counterfactual nodes cannot be reached by rational play, and thus it implies that some player must have been irrational (or must have had some doubts about the others' rationality, or must have made some error) *before* such an "irrational" node was reached; but this doesn't contradict in any way the assumption that the players *will* be rational at that node (and further in the future).

Dynamics cannot really be understood without its correlative: *invariance* under change. Certain truths, or beliefs, *stay true* when everything else changes. We have

[4] A sentence like this is called a 'Moore sentences' after G.E. Moore [1942], cf. [Segerberg, 2006; Benthem, 2004].

already encountered an absolute form of invariance: (irrevocable) knowledge, i.e. belief that is invariant under *any* possible information change. Now, we need a second, weaker form of invariance: *stability*. A truth, or a belief, is *stable* if it remains true, or continues to be believed, after any *(joint) learning of "hard" information* (via some truthful public announcement). In fact, in the case of an "ontic" (non-doxastic) fact φ, Stalnaker's favourite notion of "knowledge" of φ [Stalnaker, 1996; 2006] (a modal formalisation of Lehrer and Klein's "defeasibility theory of knowledge"), also called "safe belief" in [Baltag and Smets, 2008b], corresponds precisely to *stable belief* in φ. (But note that the two notions differ when applied to a doxastic-epistemic property, such as "rationality".) Stability can be a property of a belief or a common belief: a proposition φ is a "stable belief" if the fact that φ is belief is a stable truth, i.e. φ continues to be believed after any (joint) learning of "hard" information.

What is required for achieving the backward induction outcome is *stable belief in dynamic rationality*, either in the whole model, or at least commonly known to hold for all players. In some contexts, we can think of this condition as expressing an 'optimistic' belief-revision policy about the opponents' potential for rationality: the players "keep hoping for rationality" with respect to everybody's current and future play, despite any past irrational moves. Of course, whether or not the words "hope" and "optimism" are appropriate depends on the players' payoffs: e.g. in common interest games (in which all players' payoffs are identical at all nodes), it indeed makes sense to talk about "hoping" for opponents' rationality; while in other games, it may be more appropriate to talk about "persistent cautiousness" and a "pessimistic" revision policy.

We can now give an informal statement of our main result. In a context where there is common knowledge of open future, we will have the following.

Theorem 4.1. Dynamic rationality and common knowledge of stable belief in dynamic rationality entails the backward induction outcome.

Plausibility models for extensive games are just like plausibility models for strategic games: they are (possibly 'impure') plausibility models $(W, \preccurlyeq_i, \xi)_{i\in N}$ in which ξ associates to each state $u \in W$ an *outcome* (leaf) of the game. Our assumption of *common knowledge of open future* entails that ξ will be a surjective map, since for every outcome of the game there must be (at least) one state where it is realised. Thus, in the terminology of Chapter 1, we will in this Section only consider 'full' models; the class of (full) models for the game Γ is denoted $\mathfrak{M}_\Gamma$.

We will use a conditional belief language, with propositions for preferences (over outcomes), and propositions for outcomes (leaves). So, where we write $\mathcal{O}$ to mean $\mathcal{O}(\twoheadrightarrow)$: for every $o \in \mathcal{O}$, there is a basic proposition $\mathbf{o}$ in Ψ, and for each $i \in N$ and $\{o, o'\} \subseteq \mathcal{O}$, Ψ (the set of atomic propositions) has a proposition $\mathbf{o} <_i \mathbf{o}'$. To talk about the non-terminal nodes, we introduce the following abbreviation:

$$\mathbf{p} = \bigvee_{p \twoheadrightarrow o} \mathbf{o}\,,$$

for any $p \in \mathcal{Z} - \mathcal{O}$.

The language we have described does not have terms for extensive-form strategies, which are complex objects, and we therefore treat them as such.

If a player adopts a particular strategy, our language can encode this in terms of *the player's conditional beliefs about what she would do at each of her decision nodes*. For instance, we say that Alice "adopts the backward induction strategy" in a given state u of a model for the Centipede Game in Figure 4.4 iff the sentences $\Box_a o_1$ and $\Box_a(o_3 \mid q)$ hold at state u. Similarly, we can express the fact that Bob adopts a particular strategy, and by putting these together we can capture *strategy profiles*. A given profile is realised in a model if the correspondent sentence is true at a state of that model.

Note that, in our setting, *nothing forces the players to adopt (pure) strategies*. Recall that strategies are (sometimes needlessly) "complete" plans of action prescribing a unique choice (a belief that a particular move will be played) for each decision node of the player. But the players might simply consider all their options as equi-plausible, which essentially means that they do not have a strategy.

Examples In (any state of) model $\mathcal{M}_1$ from Figure 4.5, it is common knowledge that *both players adopt their backward induction strategies*. In contrast, in the model $\mathcal{M}_2$ from Figure 4.6, it is common knowledge that *no player has a strategy* (at any node):

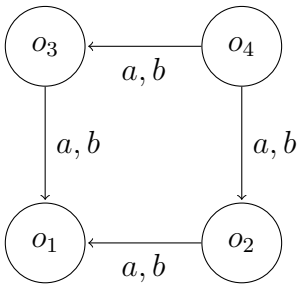


Figure 4.5: A plausibility model $\mathcal{M}_1$ for the centipede game, in which players have 'strategies'

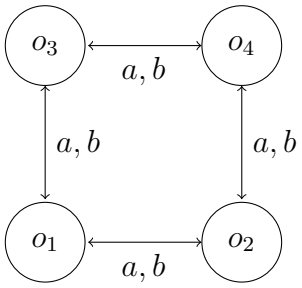


Figure 4.6: A plausibility model $\mathcal{M}_2$ in which players don't have strategies .

So the assumption that *players have strategies* is not one that we make in our models.

To think players have strategies is to make an extremely strong assumption, and one that as we have already seen contains some implicit counterfactual baggage, There is no a priori reason to assume (and there are good empirical reasons to reject) that players play according to fully-determined strategies. Our models are general enough to dispense with this assumption; indeed, our work shows that *this assumption is not needed for proving (common belief) that the backward induction strategy is played*.

We identify a player's *intentions* with *her beliefs about what she is going to do*, and so we represent the decision maker's plan of action as a belief about her (future) action. This identification is philosophically debatable, since players may be aware of the possibility of mistakes, and so they may doubt that their intentions will be realised. But one can also argue that, in a game-theoretical context, such distinctions will be of very limited significance: indeed, an intention that is not believed to be enforceable is irrelevant for strategic planning (though see [Roy, 2008] for a discussion of intentions in game theory). The players only need to know each other's beliefs about their future actions and about each other's beliefs about the others' beliefs etc., in order to make their own rational plans; whether or not they are being informed about each other's (completely unenforceable and not believed to be enforceable) "intentions" will not make any difference. So for our purposes we can safely adopt the simplifying assumption that the *players believe that they will be able to carry out their plans*. Given this assumption, a player's "intentions" can be captured by her beliefs about her (future) actions.

In the game-theoretical literature it is in effect typically assumed that, at any given moment, both the *structure of the game* and (in the case of PI games) the players' *past moves* are 'hard' information. So for example, once a move is played, all players *know, in an absolute, irrevocable sense*, that it was played: moves are "observables."

Moreover, past moves (as well as the structure of the game) are common knowledge (in the same absolute sense of knowledge). In contrast, players only really have *belief* (not knowledge) of each others' rationality, and even a player's beliefs about her own future move at some node that is not yet reached, does not attain the status of knowledge, since it has not been observed. In principle beliefs about non-observables, including one's own plans, could be revised. For instance, the player might make a mistake, failing to play according to her plan. Or the others might in fact play irrationally, forcing her to revise her belief about their rationality. So we stick to calling this kind of defeasible information 'belief'; it is based on players' soft information.

We think of every state of a game model $\mathcal{M}_\Gamma \in \mathfrak{M}_\Gamma$ as an *initial state (of a possible play)* of the game Γ. As the play goes on, the players' hard and soft information, their knowledge and beliefs, *evolve*. To represent this evolution, we will need to successively change our model, so that e.g. when a node p is reached, we want to obtain a corresponding model of the subgame Γ^p. That is precisely, in this perfect information setting, what is achieved by *updating the model with public announcements*: indeed, in a game of perfect information, every move, say from a node q to one of its immediate successors q', can be "simulated" by a public announcement $!q'$. In this way, given a model $\mathcal{M}$ of the original game Γ, then for each subgame Γ^p of Γ, we obtain a model

$\mathcal{M}^p = \mathcal{M}!\mathbf{p}$ that *correctly describes the players' knowledge and beliefs at the moment when node p is reached during a play*. Proposition 4.2 states that this is indeed a model of the corresponding subgame Γ^p.

Proposition 4.2. *If $\mathcal{M} \in \mathfrak{M}_\Gamma$ then $\mathcal{M}^p \in \mathfrak{M}_{\Gamma^p}$.*

Example Consider a play of the Centipede game that starts in the initial situation described by the model $\mathcal{M}_1$ in Figure 4.5, and in which the real state of the world is the one having outcome o_2: so Alice first plays "right", reaching node p, and then Bob plays "left", reaching the outcome o_2. The model $\mathcal{M}_1$ from Figure 4.5 gives us the initial situation, the model $\mathcal{M}_1^p$ in Figure 4.7 describes the epistemic situation after the first move, and then the model $\mathcal{M}_1^{o_2}$ in Figure 4.8 gives the epistemic situation at the end of the play:

Figure 4.7: The model $\mathcal{M}_1^p$

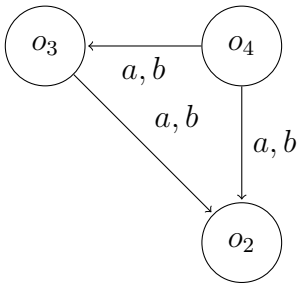


Figure 4.8: The model $\mathcal{M}_1^{o_2}$

In this way, for each given initial state s (of a given play $r, p, \ldots, o$ of the game, we obtain a *sequence of evolving game models*

$$\mathcal{M} = \mathcal{M}^r, \mathcal{M}^p, \ldots, \mathcal{M}^o\,,$$

describing *the evolving knowledge and beliefs of the players* during any play. Each model $\mathcal{M}^p$ accurately captures the players' beliefs at the moment when node p is reached. Note also that every such sequence ends with a model $\mathcal{M}^o$ in which the outcome, as well as the whole history of the game, are now common knowledge.

If we want to look at *every* subgame (in the extensive-form sense) of a given game, and to say that a given property holds *everywhere* in that game, we can use the notion of an *arbitrary public announcement*. The logic of an operator $[!]$, where $[!]\varphi$ means 'no

matter what is (truthfully) publically announced, φ will remain true', has been studied in [Balbiani *et al.*, 2008]. For our present purposes, of looking at properties that are stable in all subgames, we can also use the abbreviation:

$$[!]\varphi = \bigwedge_{p \in \mathcal{Z}} [!\mathbf{p}]\varphi.$$

We now turn our attention to defining our fundamental notions of *dynamic rationality* and *rational play*. First we will look at *single-agent (one-step) decision situations*, and then at interactive decision situations, i.e. *games*. We start with this simplest possible case in order to get a clear handle on what we have seen to be conceptually murky territory.

Given a *one-step decision problem* $\mathcal{P}$ with a set of *outcomes* $\mathcal{O}$, the *decision-maker* i selects one of the outcomes $o \in \mathcal{O}$. The decision-maker may have various hard and soft information about which outcomes can actually be realised and which not. This will determine her knowledge and her beliefs. We assume that her "hard" knowledge restricts her possible choices: she *can* only select outcomes that she doesn't know to be impossible.

What this amounts to is the following: for the decision maker i, the "true" set of possible outcomes is $\{o \in \mathcal{O} \mid \neg K_i \neg o\}$, i.e. the set of all the "epistemically possible" outcomes. So her selected option must satisfy: $o \in \{o \in \mathcal{O} \mid \neg K_i \neg o\}$. This allows us to capture the "selection" problem using epistemic operators.

To assess whether the decision is "rational" or not, one considers the decision-maker's subjective preferences, modelled as a total order $<_i$ on $\mathcal{O}$.

Rationality, in this case, corresponds to requiring that the selected option is not worse than any other (epistemically) possible alternative. In other words, i's solution of the decision problem $\mathcal{P}$ is *rational* if she does not choose any option that is strictly less preferable than an option she doesn't know to be impossible. Syntactically, we would then write something like the following formula $\mathbf{r}_i^{\mathcal{P}}$ to define the rationality of the decision-maker i in the decision problem $\mathcal{P}$.

$$\mathbf{r}_i^{\mathcal{P}} = \bigwedge_{o,o' \in \mathcal{O}} \Big((o <_i o' \wedge \neg K_i \neg o') \rightarrow \neg o \Big).$$

The main difference between our definition and the standard definition of rational decision-making is the epistemic limitation of the choice set. The epistemic operators are used here to delimit what is currently known about the availability of options: i's choice should only be compared against options that are not known to be unavailable. This is an important difference, and its importance becomes clear when we generalise our definition to extensive games, cf. the difference between 'dynamic' rationality and traditional 'substantive' rationality, described below.

We now aim to extend the above definitions to the case of multi-agent many-stage decisions, i.e. extensive games (of perfect information). Recall that in an extensive

game we are given the players' subjective preferences $<_i$ only over the leaves. However, at all the intermediate stages of the game, players have to make local choices, not between "final" outcomes, but between "intermediary" outcomes, that is: between other nodes of the game tree.

So, in order to assess players' rationality, we need to extend the subjective preference relations to all of the nodes of the game tree. Fortunately, given the above doxastic interpretation of preferences, there is an obvious (and natural) way to define these extensions. Namely, a player considers a node p to be strictly less preferable to a node q if she believes p to be strictly dominated by q. More precisely, if every outcome that she believes to be achievable given that p is reached is worse than every outcome that she believes to be achievable given that q is reached:

Definition 4.9. For each player i, we use the following abbreviation in the formal language to talk about i's ***preferences over nodes*** that are not outcomes.

$$p <_i q \;\; = \bigwedge_{o,o' \in \mathcal{O}} \Big((\neg \Box_i (\neg o | \mathbf{p}) \wedge \neg \Box_i (\neg o' | \mathbf{q})) \rightarrow o <_i o' \Big).$$

Each node $p \in \rho_i(\mathcal{Z})$ can be considered as a (distinct) decision problem, in which the decision-maker is i, the set of outcomes is the set $\mapsto (p)$ of all immediate successors of p, and the subjective preference relation is given by the (restriction of the) extended relation $<_i$ defined above. So we can define the rationality of a player i at a node $p \in \rho_i(\mathcal{Z})$ as rationality for the corresponding decision problem, i.e. the player's selection at each decision node consists only of "best answers". Note that, as before, *the player's choice is epistemically limited*: if she has "hard knowledge" that rules out some successors (for instance, because those nodes have already been bypassed), then those successors are excluded from the set of possible options. The only difference is that the "knowledge" involved is the one the player *would* have at that decision node, i.e. it is conditional on that node being reached. Formally, we obtain:

Definition 4.10. In the context of some game, the ***dynamic rationality*** of player i at node p is defined by the sentence $\mathbf{dr}_i^p$:

$$\mathbf{dr}_i^p = \bigwedge_{q,q' \hookleftarrow p} \Big((q <_i q' \wedge \neg K_i^{\mathbf{p}} \neg \mathbf{q}') \rightarrow \neg \mathbf{q} \Big),$$

Here we write $K_i^p \varphi$ for $K_i(p \rightarrow \varphi)$, in analogy with the notation $\Box_i^p \varphi$ (which, as we mentioned in the previous Chapter, we sometimes write for $B_i(\varphi | p)$), since knowledge of a conditional is the same thing as conditional knowledge (which is of course not the case for beliefs). Definition 4.10 might appear as stated not to use the belief operator, but remember that $q <_i q'$ is an abbreviation (Definition 4.9) for a sentence that *does* use player i's belief operator.

Definition 4.11. Let $\mathbf{dr}_i$ be the sentence saying that each player is rational at every node at which she plays.

$$\mathbf{dr}_i \;\; = \bigwedge_{p \in \rho_i(\mathcal{Z})} \mathbf{dr}_i^p$$

If $\mathbf{dr}_i$ is true, we say that player i satisfies *dynamic rationality*. As we'll see, asserting this sentence at a given moment is a way of saying that *the player will play rationally from that moment onwards*, i.e. she will make the best move at any *current or future* decision node.

In the following, by "dynamic rationality" we mean the formula $\mathbf{dr}$

$$\mathbf{dr} \quad = \quad \bigwedge_{i \in N} \mathbf{dr}_i$$

saying that all players are dynamically rational.

To compare our notion with Aumann's concept of "substantive rationality", we have to first adapt Aumann's definition to a belief-revision context. This has already been done by a number of authors e.g. Battigalli and Siniscalchi [1999; 2002], resulting in a definition of "rationality at a node" that differs from ours only by the absence of epistemic qualifications to the set of available options (i.e. the absence of the term $\neg K_i^p \neg q$). The notion of *substantive rationality* is then obtained from this in the same way as dynamic rationality, by quantifying over all nodes, and it is thus equivalent to the following definition.

Definition 4.12. In the context of some game, the ***substantive rationality*** of player i is defined by the formula $\mathbf{sr}_i$:

$$\mathbf{sr}_i \quad = \quad \bigwedge_{p \in \rho_i(\mathcal{Z})} \bigwedge_{q, q' \leftarrowtail p} (q <_i q' \Rightarrow \neg \mathbf{q}).$$

What is the logical connection between these two definitions of rationality?

Fact 4.2. *Substantive rationality implies dynamic rationality, i.e.*

$$\mathbf{sr}_i \Rightarrow \mathbf{dr}_i.$$

However, the converse is in general false. To better see the difference between $\mathbf{sr}_i$ and $\mathbf{dr}_i$, recall that a formula being true in a model $\mathcal{M} \in \mathfrak{M}_\Gamma$ means that it is true at the first node (the root) of the game tree Γ. However, we will later have to evaluate the formulae $\mathbf{dr}_i$ and $\mathbf{sr}_i$ at *other* nodes w, i.e. in *other models* of the form M^q (models for *subgames* Γ^q). Since *the players' knowledge and beliefs evolve during the game*, what is (not) known/believed conditional on p *in model* M^q differs from what was (not) known/believed conditional on p *in the original model* (i.e. at the outset of the game). In other words, the meaning of both dynamic rationality $\mathbf{dr}_i$ and substantive rationality $\mathbf{sr}_i$ will change during a play. But they change in different ways. At the initial node r, the two notions are equivalent. But, once a node p has been bypassed, or once the move at p has already been played by a player i, that player is counted as *rational at node* p according to our definition, while according to the usual (non-epistemically qualified) definition the player may have been irrational at p.

In other words, the epistemic limitations we imposed on our concept of dynamic rationality make it into a future-oriented concept. At any given moment, the rationality

of a player depends only on her current beliefs and knowledge, and so only on the options that she currently considers possible: past, or by-passed, options are irrelevant. Dynamic rationality simply expresses the fact that the player's decision in any future contingencies is rational (given her future options and beliefs). Unlike substantive rationality, our concept has nothing to do with the past or with contingencies that are known to be impossible: a player i may still be "rational" in our sense at a given moment/node p even when p could only have been reached if i has already made some "irrational" move. The knowledge of some past mistake may of course affect the others' *beliefs* about this player's rationality; but it doesn't directly affect her rationality, and in particular it doesn't automatically render her irrational.

So our definition of dynamic rationality makes it different from (and arguably more realistic than) Aumann's and Stalnaker's substantive rationality, but also from other similar concepts in the literature (for example Rabinowicz's [1998] "habitual" or "resilient" rationality, etc). The difference becomes more apparent if we consider the assumption that "rationality" is common belief, in the strongest possible sense, including common "strong" belief (in the sense of Battigalli and Siniscalchi [2002]), common persistent belief, or even common "knowledge" in the sense of Aumann. As persuasively argued by Stalnaker and Reny, these assumptions, if applied to the usual notions of rationality in the literature, bear no relevance for what the players would do (or believe) at the nodes that are incompatible with these assumptions! The reason is that, if these counterfactual nodes were to be reached, then by that time the belief in "rationality" would have already been publically disproved: we cannot even entertain the possibilities reachable by irrational moves except by suspending our belief (or "knowledge") in rationality. Hence, the above assumptions cannot tell us anything about the players' behaviour or rationality at such counterfactual nodes, and thus they cannot be used to argue for the plausibility of the backward induction solution (even if they logically imply it)! In contrast, our notion of dynamic rationality is *not* automatically disproved when we reach a node excluded by common belief in it: a player *may* still be rational with respect to her current and future options and decisions *even after* making an "irrational" move. Indeed, the player may have been playing irrationally in the past, or may have had a moment of temporary irrationality, or may have made some mistakes in carrying out her rational plan; but she may have recovered now and may play rationally thereafter. Since our notion of rationality is future-oriented, no information about past moves will necessarily and automatically shatter belief in rationality (although of course it *may* still shatter it, or at least weaken it). So it is perfectly consistent (although maybe not always realistic) to assume that players maintain their common belief in dynamic rationality despite all past failures of rationality. In fact, *this is our proposed solution to the BI paradox*: we will show that such a "stable" common belief in dynamic rationality (or more precisely, common *knowledge* of the stability of the players' common belief in rationality) is exactly what is needed to ensure common belief in the backward induction outcome.

It is easy to see that Aumann's theorem stating that common knowledge of substantive rationality implies the backward induction outcome [Aumann, 1995] can be

strengthened to Proposition 4.3

Proposition 4.3. *In any state of any plausibility model for a game of perfect information, common knowledge of dynamic rationality implies the backward induction outcome.*

Unfortunately, common knowledge of (either dynamic or substantive) rationality can never hold in a *full* model. It is incompatible with the condition of (epistemically) open future. By requiring that players have "hard" information about the outcome of the game, Aumann's assumption does not allow them to reason hypothetically or counterfactually about other possible outcomes, at least not in a consistent manner.[5] This undermines the intuitive rationale behind the backward induction solution.

So we must give natural conditions that can be satisfied on game models, but that still imply the backward induction outcome. Those are given in Theorem 4.1.

Theorem 4.1. *The following holds in any state u of any game model $\mathcal{M} \in \mathfrak{M}_\Gamma$:*

$$\mathbf{dr} \wedge K^*[!] \Box \mathbf{dr} \quad \rightarrow \quad BI_\Gamma,$$

where BI_Γ is the sentence $\mathbf{o}$ where o is the backward-induction outcome in the game Γ.

This is indeed a formal statement of the informal paraphrase we gave above: that dynamic rationality and common knowledge of stable belief in dynamic rationality together imply common belief in the backward-induction outcome. – The antecedent of the formula is the conjunction $\mathbf{dr} \wedge K^*[!] \Box \mathbf{dr}$; the first of these terms stands for dynamic rationality, and the second for common knowledge of stable belief in dynamic rationality.

We will now prove Theorem 4.1. First, some definitions:

Definition 4.13. For a finite set $\mathcal{O}$ of outcomes and a finite set N of players, we denote by $\mathfrak{Games}(\mathcal{O}, N)$ the class of all generic perfect information games having any subset of $\mathcal{O}$ as their set of outcomes and having any subset of N as their set of players.

Definition 4.14. For any sentence φ of our language,
φ is ***valid on a game*** Γ if φ is true at every state u of every game model $\mathcal{M} \in \mathfrak{M}_\Gamma$.
φ is ***valid over*** $\mathfrak{Games}(\mathcal{O}, N)$ if φ is valid on every game $\Gamma \in \mathfrak{Games}(\mathcal{O}, N)$.

When the game Γ is implicit from the context, we will often abbreviate BI_{Γ^p}, i.e. the name for the formula that defines the backward induction outcome in the subgame of Γ that starts at the node p, to BI^p.

[5] Indeed, if o is the backward induction outcome, then the above Proposition entails $K_i o$ for all players i, and thus for every other outcome $o' \neq o$ and every proposition φ, we have $\Box(\varphi \mid \mathbf{o}')$: the players *believe everything* (including inconsistencies) conditional on o'.

Lemma 4.1. *For every game Γ, if we denote the root of Γ by r,*

$$\left(\mathbf{dr}^r_{\rho(r)} \wedge \bigwedge_{p \leftarrow\!\!\shortmid r} \Big(\bigwedge_{q \leftarrow\!\!\shortmid r} \Box_{\rho(r)}([!\mathbf{q}]BI^{\mathbf{q}}|\mathbf{q}) \wedge [!\mathbf{p}]BI^{\mathbf{p}}\Big)\right) \;\rightarrow\; BI_\Gamma$$

is valid on Γ.

Proof. This follows directly from the definition of rationality at a node and the definition of BI. Let $i = \rho(r)$, and take any state $u \in W$ satisfying the antecedent of the claimed validity. Write $X^q_i(u)$ for $\text{MIN}_{\preccurlyeq_i}([\![\mathbf{q}]\!] \cap W^u_i)$, the set of states most plausible for u conditional on q. The assumption that $\Box_i([!]BI^q|\mathbf{q})$ is true at u means that $\xi(X^q_i(u)) = \mathfrak{o}(\text{BI})$. Since we are in a full model, $u \Vdash \mathbf{dr}^r_i$ implies that for the $p \leftarrow\!\!\shortmid r$ with $u \Vdash \mathbf{p}$, $\overrightarrow{\text{BI}}(p)$ is $<_i$-maximal for i amongst all $\overrightarrow{\text{BI}}(q)$ with $q \leftarrow\!\!\shortmid r$. But that means that this p is the one chosen by the backward induction algorithm. Given this backward-induction choice (p) of i at node r, and given the fact (ensured by the condition $[!p]BI^p$) that starting from node p everybody *will* play the backward induction choices, we can conclude that the outcome $\xi(u)$ belongs to the backward induction set of outcomes for the game Γ. Hence u satisfies BI_Γ. ■

The main Lemma underlying our result is the following:

Lemma 4.2. ***("Main Lemma")** Fix a finite set $\mathcal{O}$ of outcomes and a finite set N of players. Let φ be any sentence such that for every game $\Gamma \in \mathfrak{Games}(\mathcal{O}, N)$ with root r the following is valid on Γ:*

$$\varphi \rightarrow \left(\mathbf{dr}^r_{\rho(r)} \wedge \bigwedge_{q \leftarrow\!\!\shortmid r} \Box^{\mathbf{q}}_{\rho(r)}[!\mathbf{q}]\varphi \wedge \bigwedge_{p \leftarrow\!\!\shortmid r} [!\mathbf{p}]\varphi\right)$$

Then we have that

$$\varphi \Rightarrow BI_\Gamma$$

is valid over $\mathfrak{Games}(\mathcal{O}, N)$.

Proof. We need to prove that, for every game $\Gamma \in \mathfrak{Games}(\mathcal{O}, N)$, the sentence $\varphi \Rightarrow BI_\Gamma$ is valid on Γ. The proof is by induction on the length of the game Γ.

For games of length 0 (only one outcome, no available moves), the claim is trivial (since the only possible outcome is by definition the backward induction outcome).

Then let Γ be a game of length $n > 0$, and assume the claim is true for all games of length $< n$. Let r be the root of Γ, $i = \rho(r)$, $\mathcal{M} \in \mathfrak{M}_\Gamma$ be a model of Γ, and u be a state in $\mathcal{M}$ such that $u \Vdash \varphi$.

Take $q \leftarrow\!\!\shortmid r$. By the property assumed in the statement of this Lemma, we have $\mathcal{M}, u \Vdash \Box^{\mathbf{q}}_i[!\mathbf{q}]\varphi$, and so (again letting $X^q_i(u) = \text{MIN}_{\preccurlyeq_i}([\![\mathbf{q}]\!] \cap W^u_i)$), then we have $\mathcal{M}, v \Vdash [!\mathbf{q}]\varphi$ for all $v \in X^q_i(u)$. Hence, we have $\mathcal{M}^q, v \Vdash \varphi$ for all $v \in X^q_i(u) \cap [\![\mathbf{q}]\!]$. By the inductive hypothesis, $\mathcal{M}^q, v \Vdash \text{BI}^q$ for all such v. Therefore $\mathcal{M}, v \Vdash [!\mathbf{q}]BI^u$ for all $t \in X^q_i(u)$, and hence that $\mathcal{M}, u \Vdash \Box^{\mathbf{q}}_i[!\mathbf{q}]BI^q$.

Now let $p \leftarrow\!\shortmid\ r$ be such that $u \Vdash p$. By the property assumed in this Lemma, we have that $\mathcal{M}, u \Vdash [!\mathbf{p}]\varphi$. By the same argument as in the last paragraph, the inductive hypothesis gives us $\mathcal{M}, u \Vdash [!\mathbf{p}]\mathrm{BI}^p$. Putting together with the conclusion of the last paragraph and with the fact (following from the Lemma's hypothesis) that $\varphi \Rightarrow \mathbf{dr}_i^r$ is valid on $\mathcal{M}$, we infer that $\mathcal{M}, u \Vdash \mathbf{dr}_i^r \wedge \bigwedge_{q \leftarrow\!\shortmid\ r} \Box_i^{\mathbf{q}}[!\mathbf{q}]\mathrm{BI}^q \wedge \bigwedge_{p \leftarrow\!\shortmid\ r}[!\mathbf{p}]BI^p$. The desired conclusion follows now from Lemma 4.1. ■

Lemma 4.3. *The sentence*

$$\mathbf{dr} \wedge K^*[!] \Box\, \mathbf{dr}$$

has the property assumed in the statement of Lemma 4.2.

Proof. The claim follows from the following three sub-claims.

1. Dynamic rationality is a "stable" property,
   i.e. the implication $\mathbf{dr} \rightarrow [!]\mathbf{dr}$ is valid.
2. The implication $K^*[!] \Box\, \psi \rightarrow \Box_i^{\mathbf{q}}[!\mathbf{q}]K^*[!] \Box\, \psi$ is valid,
   for all formulae ψ and all nodes $q \in \mathcal{Z}$.
3. The implication $K^*[!] \Box\, \psi \rightarrow [!\mathbf{q}]K^*[!] \Box\, \psi$ is valid,
   for all formulae ψ and all nodes $q \in \mathcal{Z}$.

■

Theorem 4.1 follows now from Lemma 4.2 and Lemma 4.3. Another sentence with the property in Lemma 4.2 is given in [Baltag *et al.*, 2009], where the notion of *stability* of belief is investigated further, and 'stable true belief' is introduced and studied, in particular being used to formulated an alternative condition for backward induction.

Subgame-perfect equilibrium is a refinement of Nash equilibrium, and so is properly speaking an equilibrium notion. The epistemic 'explanation' that would usually be given for it would therefore follow the steady-state interpretation of game theory, saying that players have correct beliefs about each others' strategies, and are rational.

However, since in generic games there is a unique subgame-perfect equilibrium, there is an explanation based on *deductive* notions as to why players would play according to the subgame-perfect equilibrium outcome. We gave those in Theorem 4.1; but we also have a stronger result.

Corollary 4.1. *The following holds in any state s of any model $\mathcal{M} \in \mathfrak{M}_\Gamma$:*

$$K^*[!] \Box\, \mathbf{dr} \rightarrow K^*[!] \Box^*\, BI_\Gamma$$

Proof. Follows from Theorem 4.1, by applying the operator $K^*[!]\Box$ to both its premise and its conclusion, and noting that the following implication is valid:

$$K^*[!] \Box\, \psi \rightarrow K^*[!] \Box\, K^*[!] \Box\, \psi.$$

■

Corollary 4.1 means that the same conditions as those for Theorem 4.1 entail also that players have common knowledge of *each other's strategies as beliefs*. So a purely deductive approach leads to a situation where there is common belief among the players of their strategies.

4.3 Games with imperfect information

Trembling-hand equilibrium was also introduced by Selten [1975], as a refinement of his earlier concept of subgame-perfect equilibrium. Intuitively, the point about trembling-hand equilibrium is that it takes seriously the idea that players might make mistakes, and in effect integrates this idea formally into the definition of the solution concept. Trembling-hand equilibrium is a *refinement* of subgame-perfect introduction, so a trembling-hand equilibrium is also a subgame-perfect equilibrium. In generic games of perfect information, like the ones we have looked at so far, it is known that subgame-perfect equilibria are also trembling-hand equilibria ([Fudenberg and Tirole, 1991, Section 8]), meaning that on generic perfect-information games, trembling-hand equilibrium is equivalent to subgame-perfect equilibrium.

However, in a wider class of games, the equivalence does not hold. Trembling-hand equilibrium was originally formulated for games of so-called 'imperfect information'. Games of imperfect information, defined below (see Definition 4.15) are those in which players do not always collectively publically observe each other's moves; so they make choices not from (the point of view of) a node in the tree, but from a so-called 'information set', that is something like a relational model on the tree. In any case, the natural extension of the notion of subgame-perfect equilibrium for extensive-form games, arrived at by a slight re-definition of the notion of a subgame. And according to this definition, trembling-hand equilibrium is a *strict* refinement of subgame-perfect equilibrium with respect to extensive games of *imperfect* information, meaning that there are extensive games of imperfect information with subgame-perfect equilibria that are *not* trembling-hand equilibria.

Indeed, Selten [1975] motivated trembling-hand equilibrium using an example of an extensive game (with imperfect information), known as 'Selten's Horse', in which there is a strategy that is a subgame-perfect equilibrium but intuitively should not be played, by the same sort of reasoning that led to the notion of subgame-perfect equilibrium.

From our dynamic epistemic logic perspective, the information flow in games of imperfect information can be modelled by considering not public announcements, but more complex epistemic actions. So while any action that is a move in a game of perfect information can be thought of as a public announcement, an action in a game of imperfect information is a *private announcement*: some players learn what has happened, the other players learn that those players have learnt what has happened, without themselves learning what move happened. We will explain how this works below. As we saw in Section 3.1, DEL provides the facility for many more kinds of epistemic

action than just private announcements, and extensive games of imperfect information, as they are currently defined, only exploit relatively primitive forms of uncertainty.

We will remark that an obvious variation of the definition of subgame-perfect equilibrium, already indicated in [Osborne and Rubinstein, 1994], takes care of the particular 'Horse' example that Selten used to motivate his concept of subgame-perfect equilibrium. That leads us to remark that that solution concept, which is a refinement of the traditional subgame-perfect equilibrium, can be grounded in terms of dynamic epistemic logic.

Trembling-hand equilibrium is defined in terms of limits of a sequence of totally mixed strategy profiles, and so is only defined in terms of games of cardinal preferences. We introduce a further *refinement* of trembling-hand equilibrium, that we call 'even-handed trembling-hand equilibrium'. It boils down to saying that all players are equally (infinitesimally) likely to deviate. We then look at how to define a version of it that applies also to games with ordinal preferences, giving a *definition of the solution concept in terms of its epistemic conditions*, that can be see as a combination of our notion of lexicographic rationality along with a specific belief revision policy.

Recall that a game of perfect information was a tuple of the form

$$(\mathcal{Z}, \twoheadrightarrow, \rho, \leq_i)_{i \in N}.$$

In games of *imperfect* information, we add a component to capture the fact that some moves might not be observed by all players. This is defined by introducing an indistinguishability relation $\mathcal{I}_i$ for each player. Now, the only time it *matters* whether i knows where she is in the game is when it is i's turn. Therefore the relation runs over $\rho_i(\mathcal{Z})$. Thus, for each $i \in N$, we let $\mathcal{I}_i$ be an *equivalence* relation (an S5 relation) on the set $\rho_i(\mathcal{Z})$. We also write $\mathcal{I}_i$ to mean the induced partition:

$$\{X \subseteq \rho_i(\mathcal{Z}) \mid \forall x, y \in X,\ x \mathcal{I}_i y, \text{ and } \forall x \in X\, \forall z \notin X,\ \neg(x \mathcal{I}_i(y))\}$$

The elements of $\mathcal{I}_i$ are called i's '***information sets***'. We use the usual lifting of relations to functions, so that for any node $p \in \rho_i(\mathcal{Z})$, we write $\mathcal{I}_i(p)$ to mean the (unique) information set $I \in \mathcal{I}_i$ such that $p \in I$. We need to place one restriction on this information partition: if i is expected not to be able to distinguish between p and q, then because i has to make a choice at whichever node is reached, i *must have the 'same' options available* at p and q (since as always the game is assumed to be known to i). This is motivated by the following line of argument: suppose that at p, i can choose between two options, L and R, but at q has only one option. Then if i is at p, she can reason as follows: There are two options, whereas at q there would only be one option. Therefore we are not at q. We must stipulate just that if i cannot tell the difference between p and q, then p and q must have the same number of successors:

$$p \mathcal{I}_i q \;\Rightarrow\; \#(\mapsto (p)) = \#(\mapsto (q)).$$

However, in games of imperfect information, we will actually include the action labels L in the definition of a game, because we want also to be able to talk about a player being faced with the 'same' choices at two different nodes, and not just the same *number*

of choices. Then where ℓ is such a labelling function, the condition that we will impose on the relation $\mathcal{I}_i$ for each player $i \in N$ is the following condition, that is strictly stronger that the previous condition in terms of cardinality.

$$p\mathcal{I}_i q \Rightarrow \ell(\mapsto (p)) = \ell(\mapsto (q)).$$

That is: the actions available to i, in any two nodes p and q that are indistinguishable for i, are the same, because otherwise i would be able to distinguish between p and q. Note that now we will need to impose the condition on naming functions that *no two successors are assigned the same name*, i.e.

$$p \mapsto q \;\&\; p \mapsto q' \;\; \Rightarrow \;\; \ell(q) \neq \ell(q')$$

Definition 4.15. A ***game (in extensive form) with imperfect information*** is a tuple

$$(\mathcal{Z}, \ell, \twoheadrightarrow, \rho, \mathcal{I}_i, \leq_i)_{i \in N},$$

where the components are as described above.

There is another standard condition that it makes sense to impose on player i's information relation/partition $\mathcal{I}_i$, and that is a condition of *perfect recall*. In these games, the idea is that players do not forget information, so that if player i knew at some stage that node p was reached, and p is incompatible with q being reached, then she will not think that q has been reached. Later in this Chapter we show how to extend the information partition on an extensive game to the full game tree, in a way that will require this condition of perfect recall. In perfect recall games, players do not 'forget' their information sets, and also do not forget their own moves. So define $X_i(p)$ recursively by the distance of p from the root r. If $p = r$ then set: $X_i(p) = \{\emptyset\}$. Otherwise, suppose that for the predecessor q of p (i.e. $p \mapsto q$), $X_i(q)$ is defined. What we want is to have X_i record any moves that i has just made, and any information partition she finds herself in. So if $\rho(p) = i$, let $Y = \{\mathcal{I}_i(p)\}$, otherwise $Y = \emptyset$; Y gives the information partition that i finds herself in, if any. and if $\rho(p) = i$ then $Z = \{\ell(p)\}$, otherwise $Z = \emptyset$; Z gives the move i has just made, if any. So then let $X_i(p) = X_i(q) \cup Y \cup Z$, and say that an extensive game has *perfect recall* just if, when $q\mathcal{I}_i q'$ we also have $X_i(q) = X_i(q')$.

In the context of a game of imperfect information, a ***strategy*** for i is a function from i's the elements of i's *information partition* to the *labels* of the successors of that element, i.e. s_i is a strategy if it's of the form

$$\begin{array}{rccl} s_i: & \mathcal{I}_i & \to & L \\ & I_i & \mapsto & s_i(I_i) \in \ell(\mapsto (I_i)) \end{array}$$

Then in order to define, as in the simpler case of games of perfect information, the *outcome* $\mathfrak{o}(s)$ of a strategy profile s, let f_s be the function from nodes to nodes that

is induced in the natural way by a given strategy profile s. That is, the function that associates, to every node p the successor $q \leftarrow\!\!\shortmid p$ determined by the label that in turn is determined by the strategy profile and the information set to which p belongs. In more formal notation:

$$\begin{array}{rccl} f_s & \mathcal{Z} & \to & \mathcal{Z} \\ & p & \mapsto & \text{the } q \leftarrow\!\!\shortmid p \text{ such that } \ell(q) = s(\mathcal{I}_{\rho(p)}(p)) \end{array}$$

Then we define $\mathfrak{o}(s)$ as the unique leaf such that there is a natural number $k \in \mathbb{N}$ with $f_s^k(r)$, where r is the root of the game tree.

Now we can define the normal form of an extensive game of imperfect information in the same way as we defined the normal form in the perfect information case. Furthermore with imperfect information it's now also possible to give, for any normal form game G an extensive form game $\Gamma(G)$ that is equivalent to it: simply have one information set for each player i, occurring in any order, and with each one branching $\#(T_i)$ times.

A subgame of an extensive game with imperfect information must respect the information partition.

Definition 4.16. A ***subgame*** of

$$(\mathcal{Z}, L, \twoheadrightarrow, \rho, \mathcal{I}_i, \leq_i)_{i \in N}$$

is any tuple

$$(\mathcal{Z}', L\restriction\mathcal{Z}', \twoheadrightarrow \restriction\mathcal{Z}', \rho\restriction\mathcal{Z}', \mathcal{I}_i\restriction\mathcal{Z}', \leq_i \restriction\mathcal{Z}')_{i \in N}$$

such that $(\mathcal{Z}', L\restriction\mathcal{Z}')$ is a *tree* and for any $p \in \mathcal{Z}'$ and $q \in \mathcal{Z}$ with $p\mathcal{I}_i q$, we have $q \in \mathcal{Z}'$.

The way in which subgame-perfect equilibrium is then defined for games of *imperfect information*, for example in [Selten, 1975], is effectively: that s is a subgame-perfect equilibrium in Γ if, for every subgame Γ' of Γ, in the normal form of Γ', s restricted to Γ' is a Nash equilibrium.

Figure 4.9 depicts an extensive game with imperfect information that is used as an example by Selten [1975] in order to motivate a refinement to his concept of *subgame-perfect equilibrium*, which, as we have mentioned, he argues is too permissive. Selten illustrates using this "Horse"[6] example that there are subgame-perfect strategies that nonetheless are not *intuitively* rational to play in equilibrium, where, as we would say, all beliefs would be common beliefs. He then introduces a refinement of subgame-perfect equilibrium, which he calls simply "perfect equilibrium", now more commonly known as "trembling-hand equilibrium".

Each of the three players a, b, c has just one information partition, so we can denote a strategy profile by a triple (ℓ_a, ℓ_b, ℓ_c), where ℓ_i denotes the label of a's chosen action.

[6] Selten himself refers to the example as his "numerical example"; the more colourful and now standard moniker came later.

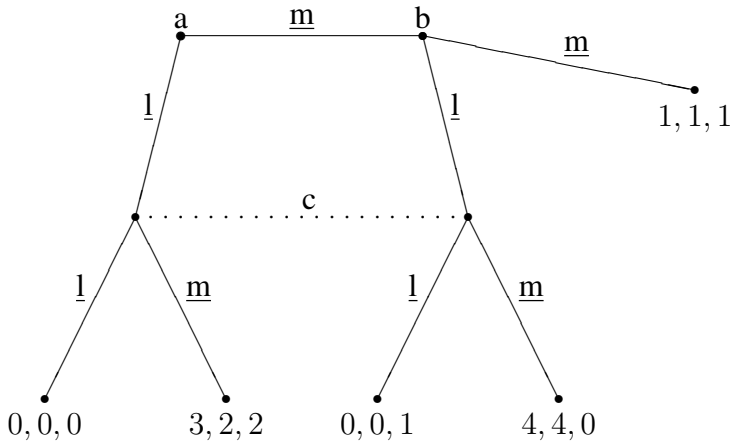


Figure 4.9: Selten's Horse

Selten points out that there are two Nash equilibria[7]:

s. (m, m, l)

s'. (l, m, m)

And since *this game has no subgames*, each of these equilibria is immediately subgame perfect.

However, similar reasoning to that motivating the concept of subgame-perfect equilibrium applies here, to rule out s' as a reasonable steady state. What exactly is intuitively objectionable about this? The problem lies in the unrealised intention of player b. Intuitively he is not an equilibrium choice: if b were actually allowed to exercise his choice, while still believing in the *rest* of the equilibrium, then b's choice in s' is not rational, since it not a best response to his anticipation about what c will do: his actual choice would (according to his equilibrium (steady-state) expectation of c's choice) give him 1, whereas he has an alternative that would (according to that same expectation) yield 4.

Of course, if the game were one of perfect information, in which c were able to distinguish between both nodes in her information set, then the node at which b makes a decision would define a subgame. And there is a sense in which for everybody but player c the game *is* a game of perfect information. A more natural analogue of subgame-perfect equilibrium, that call 'subtree-perfect equilibrium' is the following:

Definition 4.17. s is ***subtree-perfect*** just for each player i, s_i is optimal *at every information set of* i if the other players would then play according to s_{-i} (cf. [Osborne and Rubinstein, 1994, p. 219])

[7]Selten shows actually that there are two kinds of *mixed* Nash equilibria, and considers mixed strategies in extensive games. We have no need to introduce these and so talk just about pure strategies; s and s' are the pure equilibria among the mixed ones.

We can use the epistemic actions of DEL in order to ground the solution concept of subtree-perfection. However, in the more general case of *imperfect information*, this simple relativisation is not sufficient to capture the epistemic subtleties involved. That is because a move like player a choosing R in Selten's horse is *public for* a *and* b but *hidden from* c.

To flesh out the connection, we will associate to each game Γ an *action model* $\mathcal{A}_\Gamma$ containing an epistemic action e_p for every node p in Γ. The idea will be that when e_p is applied to an initial *state* model $\mathcal{M}_\Gamma$ of the game Γ, that model will change in precisely the way it *should* change to represent the epistemic effects of p being played. This will sometimes *not* be a model of any subgame of Γ; for example in Figure 4.9, the model specifying the epistemic situation that arises if player b has the chance to actually make a move will not be a model of any subgame.

Thus while $\mathcal{M}_\Gamma$ represents the beliefs of the players before play has started, $\mathcal{M} \otimes e_p$ will represent the beliefs the players would have if node p of the game were to be reached.

The precondition for the action e_p is just the disjunction of outcomes with which p is compatible:

$$\text{PRE}(e_p) \;=\; \bigvee_{o \leftarrow p} \mathbf{o}$$

In order to define, given some game Γ, the uncertainty relations $\rightarrow_i$ in the action model $\mathcal{A}_\Gamma$, we essentially just need to *extend* the existing information partitions of the players, which only run over *their own* nodes, to a partition of *the entire game tree*.

So in Selten's Horse for example (Figure 4.9), what should we say about c's beliefs before a has played? Or indeed when b is about to play? Or when b has played? The intuition is to extend the notion of "perfect recall", so that for example at the node where b makes a choice, c should not be aware of what move a has made (i.e. that b is even about to make a move), since potentially later along the same path c will be uncertain what a's move was. Yet clearly if she could recall b making a choice, then she would be able to work out what a's move was, and so (by perfect recall) would not be uncertain as to what node she was in at her information set. Similarly, she could work out that since if b had had the opportunity to move then she (c) would have been aware of it, then if a plays down she would realise that had happened too.

We will define a unique extension of the given uncertainty in this way. Note that a number of alternative versions are available, which for games with more nodes and more uncertainties would give differing models, and a coherent story could be given for all of them. An imperfect information game, with only information sets for i of nodes at which i plays, is supposed to be a sufficient basis for epistemic analysis, and for reasoning about the players' moves in the game. Therefore all of these different ways of extending the uncertainty should yield, given the same game, action models that, while different, are the same in all significant respects when it comes to reasoning about the players' moves in the game. So our canonical choice should be without loss of generality.

Consider the game in Figure 4.10. Suppose there that player 1 plays left. Does

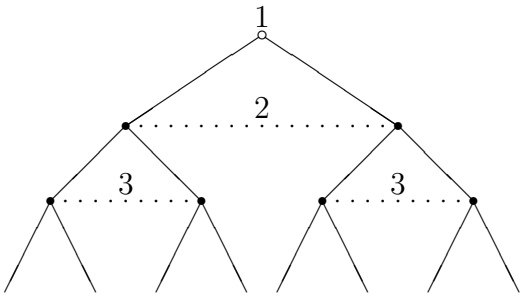


Figure 4.10: An extensive game where the players' information set does not always determine what each knows when

player 3 think player 1 has moved, but is unsure what move he has made? Our point here is simply that the game does not specify this, since for existing analyses it is unimportant. Of course, if an outcome is reached, we might want to say that all players are informed of it, so for example player 2's uncertainty would be resolved. Again, this is not in the model, and so it cannot matter for the game-theoretical analysis what choice we make, but we will be telling a slightly different 'story' of what (social-)epistemic situation an extensive game represents.

The choice we make is to "maximise past uncertainty." That is, we assume that while the player does not 'forget' anything, she does not 'learn' anything either, until the game explicitly states that she does. So given an information set I_i for player i, we move backwards one step from this information set, to the set $\leftarrowtail (I_i)$ $(= \{p \in \mathcal{Z} \mid \exists q \in I_i : p \mapsto q\})$. Intuitively all of these should be indistinguishable for i (note that $\leftarrowtail (I_i)$ can be a singleton even if I_i is not). We want to keep tracing back and identifying *all of the nodes we pass*, for i, until we reach nodes that i *can* distinguish. And, since we are in a game with perfect recall, we can say this only happens when i has just played.

More formally, let $Z_i(I) = \{p \in \leftarrowtail (I) \mid \neg\exists q \mapsto p : q \in \rho_i(\mathcal{Z})\}$. Then since our trees are finite, there is some m_I^i such that

$$\bigcup_{l<m_I^i} Z_i^l(I) = \bigcup_{l<m_I^i+1} Z_i^l(I).$$

Then for each $p \in \bigcup_{l<m_I^i} Z_i^l(I)$, we define the relation $e_p \rightarrowtriangle_i e_q$ iff $q \in \bigcup_{l<m_I^i} Z_i^l(I)$.

In this way we can work through the entire tree and define, for each node p, and each player i, for which nodes q we should have $e_p \rightarrowtriangle_i e_q$. This defines an action model $\mathcal{A}_\Gamma$ that, along the lines of the BM thesis, represents all of the different actions that are possible in the game Γ, so that given some model $\mathcal{M}$ representing the players' initial beliefs before the game (i.e. at p, the root node), the new model $\mathcal{M} \otimes \mathcal{A}_\Gamma$ is a model in which every possible action in the game has occurred. That means that if the

actual state is u, then the state (u, e_p) in $\mathcal{M} \otimes \mathcal{A}_\Gamma$ represents the epistemic situation of the players after they have moved to p.

We interpret this as backing up our claim that extensive games, even of imperfect information, do not have many informational subtleties to them: we have shown that it is straightforward to represent the information flow within an extensive game via a DEL action model. In particular, there are richer actions, as well as *soft information*, none of which are currently used in extensive games.

We should turn now to defining trembling-hand equilibrium, which we define with respect to strategic games. The definition of trembling-hand equilibrium for extensive-form games is equivalent to saying that a strategy profile is a trembling-hand equilibrium in the so-called "agent normal form" of the game [Selten, 1975]. In the agent normal form of Γ, one includes one player for each *information set* in Γ. The idea is that players' mistakes should be viewed as independent: if you tremble now it does not mean that you will tremble later. Therefore in a sense you are a different '*agent*' at the different nodes. Still, each of these agents is in a certain sense still the same *player*: they all have the same 'exogenous objectives': the same preferences over the outcomes.

Trembling-hand equilibrium is defined in terms of *totally mixed* strategy profiles.

Definition 4.18. A mixed strategy σ_i of player i is ***totally mixed*** just if for every $s_i \in T_i$, $\sigma_i(s_i) \neq 0$.

Clearly then, we will be concerned here with games with *cardinal preferences*, and it is not unproblematic to extend the definition to the infinite case, so we can consider only finite games. A totally mixed strategy profile is just a profile of totally mixed strategies.

Definition 4.19. A strategy profile σ is a ***trembling-hand equilibrium*** just if there exists a sequence $(\sigma^m)_{m\in\mathbb{N}}$ of totally mixed strategy profiles that converges to σ, and such that for each player i, σ_i is a best response against all σ^m_{-i}, i.e.:

$$\forall \sigma'_i \in \Delta T_i,\ \mu_i(\sigma_i, \sigma^m_{-i}) \quad \geq \quad \mu_i(\sigma'_i, \sigma^m_{-i})$$

Halpern [2008] has provided an elegant characterisation of a number of refinements of Nash equilibrium, including trembling-hand equilibrium, in terms of non-standard probabilities.[8] Our main concern is not to give an exposition of trembling-hand equilibrium, but rather to present a refinement of it that makes sense from the point of view of games with ordinal preferences, and which is at least a little bit closer to the deductive interpretation. So we will only quickly present Halpern's characterisation of trembling-hand equilibrium; we refer to [Halpern, 2008] for more detail. let ε be an *infinitesimal*, i.e. an entity smaller than any real number yet greater than zero. If we allow ε into our definition of a mixed strategy, we get *non-standard mixed strategies*. So for example, over $T_i = \{W, M, D\}$, the following would be a non-standard mixed

[8] See [Keisler, 2000] for an introductory textbook on non-standard analysis.

strategy: $\{(W, 0.3+\varepsilon), (M, 0.6-3\varepsilon+\varepsilon^2), (D, 0.1+2\varepsilon-\varepsilon^2)\}$. Given a non-standard (mixed) strategy σ_i^ε, we say that the strategy σ_i *differs infinitesimally from* σ_i^ε just if, should ε be evaluated as 0, then the two would agree on the probability of all pure strategies $s_i \in T_i$. Then essentially what Halpern shows is the following:

Theorem 4.2 (Cf. [Halpern, 2008, Theorem 1.4])**.** *σ is a trembling-hand equilibrium just if there is a totally mixed (possibly nonstandard) strategy profile σ^ε such that for each player i, σ_i^ε differs infinitesimally from σ_i, and σ_i is a best response against σ_{-i}^ε.*

We can think of the nonstandard strategy profile as giving, in some sense, the *conditional beliefs* of the players: in an equilibrium σ, the players do all believe that they play according to σ, but *in case of ties*, they will also consider the 'fallback' possibilities, in the order of importance that is determined by the non-standard strategy profile.

We would like to simplify this concept. Our simplification makes it non-numerical and so applicable to games with ordinal utilities, and at the same time more intuitively accessible, so that we are able to give an epistemic foundation for it, along the lines of the simple epistemic characterisation of Nash equilibrium that we saw above.

Rather than allowing an arbitrary nonstandard probability distribution, we could suppose instead that *all trembles* (deviations) *are equally likely*. There are two ways to cash this out. The first is to suppose that for any strategy s_j' not played in equilibrium, s_j' occurs with the infinitesimal probability ε. This entails that players with more strategies are (infinitesimally) more likely to tremble. The second is to suppose that each player has an equal ('infinitesimal') probability ε of deviating from the equilibrium strategy profile σ, all such trembles of that player are then equally likely. We choose the first option, but will see that the two do not lead to the same set of best responses,

Let us then define even-handed trembling-hand equilibrium in the framework of Halpern's [2008] characterisation of trembling-hand equilibrium. Again assume a finite game, and write $\#i$ for $\#(T_i) - 1$, the number of strategies i could play other than whatever equilibrium strategy she actually does play. We will eventually be interested in games with *ordinal* preferences, so let σ be a *pure* strategy profile.

Then we define a specific nonstandard totally mixed strategy profile σ^e, by setting, for all $i \in N$, σ_i^* as the following nonstandard totally mixed strategy:

$$\sigma^*(s_i) \quad = \quad \begin{cases} 1 - \#i.\varepsilon & \text{if } s_i = \sigma_i \\ \varepsilon & \text{otherwise} \end{cases}$$

Clearly each σ_i^* is a non-standard probability distribution over T_i, i.e. a non-standard mixed strategy for i; furthermore σ_i^* is *totally mixed*, and differs only infinitesimally from σ_i.

Definition 4.20. σ is an ***even-handed trembling-hand equilibrium*** just if for each player $i \in N$, σ_i is a best response to $\sigma*_{-i}$.

The second strategy profile, that we could have used to define even-handed trembling hand equilibrium, is very similar: for all $i \in N$,

$$\sigma^x(s_i) \;=\; \begin{cases} 1-\varepsilon & \text{if } s_i = \sigma_i \\ \frac{\varepsilon}{\#i} & \text{otherwise} \end{cases}$$

As the game in Figure 4.11 demonstrates, these two totally mixed non-standard strategy profiles σ^* and σ^x do not yield the same best responses, and so the definition of even-handed trembling-hand equilibrium would not be the same if we used σ^x in place of σ^*. In Figure 4.11, we show the strategic game as an equivalent game of imperfect

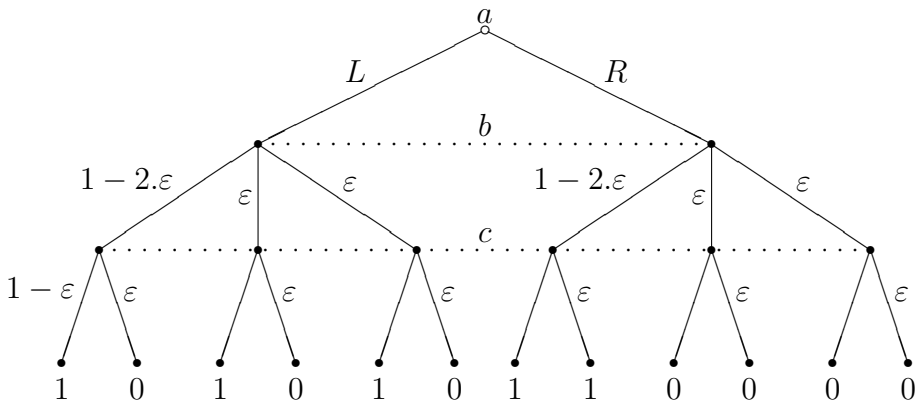


Figure 4.11: A game that shows the difference between σ^* and σ^x

information. We only give the payoffs for player a, who is choosing between L and R; we assume that the payoffs for the other players are such that whether or not a plays L or R, they play their left-most strategy in equilibrium.

Both L and R are *Nash* equilibria for player a, since they are both best responses to player a's *un*conditional beliefs in equilibrium. Let σ_{-a} denote the opponents' (b and c's) strategy profile where they both play their left-most option. Then notice that R is *not* a best response to σ^*_{-a}, since

$$\begin{aligned} \mu_i(L, \sigma^*_{-a}) &= (1-2.\varepsilon)(1-\varepsilon) + 2.\varepsilon(1-\varepsilon) \\ &> \varepsilon(1-2.\varepsilon) + (1-2.\varepsilon)(1-\varepsilon) \\ &= \mu_i(R, \sigma^*_{-a}). \end{aligned}$$

Therefore R is *not* an even-handed trembling-hand equilibrium. However, it *is* a best-response to σ^x_{-a}, since

$$\begin{aligned} \mu_i(L, \sigma^x_{-a}) &= (1-\varepsilon)(1-\varepsilon) + 2.\frac{\varepsilon}{2}(1-\varepsilon) \\ &= \varepsilon(1-\varepsilon) + (1-\varepsilon)(1-\varepsilon) \\ &= \mu_i(R, \sigma^x_{-a}). \end{aligned}$$

Those subtleties lie only in the outlying regions of our main focus here however. What we are really interested in is taking the fundamental notion behind trembling-hand equilibrium and finding a non-numerical analogue to it.

What we therefore do in what remains of this Chapter is the following: we will describe a plausibility ordering over the states of a model, that captures the intuition behind trembling-hand equilibrium, though it is closer to even-handed trembling-hand equilibrium. This ordering essentially states that players think it most plausible that they will play according to the equilibrium prediction, but that *if they do not* then *one* deviation from that prediction is the most likely; otherwise *two* deviations, and so on.

We now assume we are working with an N-player ordinal-preference game $(T, <)$. Given two (pure) strategy profiles σ and σ', define $\delta(\sigma, \sigma')$ as the number of strategies on which σ and σ' disagree. That is:

$$\delta(\sigma, \sigma^*) \;=\; \#\{i \in N \mid \sigma_i \neq \sigma_i^*\}.$$

This 'distance' function can be used to place a constraint on plausibility orderings. Recall that if players play a Nash equilibrium then they believe that the equilibrium strategy is being played. The constraint we introduce generalises that idea, so it is parametrised by a particular (pure) strategy profile σ, and we will call the constraint *respecting* δ_σ. It says roughly that for any σ', σ'' with $\delta(\sigma, \sigma') < \delta(\sigma, \sigma'')$, the ordering has at least one state u where σ' is played, and such that *there is no state as plausible as* u at which σ'' is played. That is:

$$\begin{aligned}&\forall w \in W,\, \forall \sigma', \sigma'' \in T,\, \delta(\sigma, \sigma') < \delta(\sigma, \sigma'') \;\Rightarrow\\ &\quad \exists u \in W_i^w :\, \xi(u) = \sigma' \text{ and } \forall v \in W_i^w,\, \xi(v) = \sigma'' \Rightarrow u \prec_i v\end{aligned}$$

Suppose we have a model where all players' preferences respect δ_σ. Then what are the consequences of the players being *lexicographically rational*? (Where we take rationality to be with respect to avoiding weakly dominated strategies.) Fact 4.3 says that the induced solution concept is a refinement of Nash equilibrium.

Fact 4.3. *If $\forall i \in N$, $\preccurlyeq_i$ respects δ_σ, then at all states where players are and play σ, σ is a Nash equilibrium.*

Proof. Respecting δ_σ entails being 'centred' on σ, so that $\xi(\text{MIN}_{\preccurlyeq_i}(W)) = \sigma$. And since being lexicographically rational entails being rational tout court, which entails playing a best response to your (unconditional) beliefs, then σ is a Nash equilibrium. ■

Let us define a strategy profile σ in a game of *ordinal* preferences as an even-handed trembling-hand equilibrium just when it is playable under lexicographic rationality in some model where players' plausibilities respect δ_σ. Then we have defined a solution concept in terms of a plausibility ordering. To see it another way, and to revert to the belief revision terminology that underlies the justification for using plausibility orderings to represent (rational) beliefs: we have defined a solution concept in terms

of a *belief revision policy*. The conditional beliefs 'pre-encode' [Benthem, 2007a] a player's tendency to change her beliefs.

Different belief revision policies, in concert with our notion of lexicographic rationality, will induce other solution concepts.

In this discussion of extensive-form games and trembling-hand equilibrium we moved to the strategic form of the game. Our reduction of strategies in extensive-form games, to beliefs, in concert with other aspects of our analysis of perfect-information games, unfortunately means that we cannot extend it in an elegant way to extensive games with imperfect information. Let us take a moment to mention one fascinating example of extensive-form games that we would hope would be tractable in some kind of dynamic epistemic framework, but for which we currently have no adequate proposal.

That example is what is called 'strategic communication', and the key example game is known in the literature as 'battle of the sexes with an outside option' [Osborne and Rubinstein, 1994, Figure 110.1]. This example is built by adding to the game known as battle of the sexes, which is a two-player coordination game (depicted in Figure 4.12 below, cf. Figure 2 from the Introduction) between Alice and Bob in which Alice prefers one of the options on which they could coordinate, and Bob the other. Here we give the game with cardinal preferences that stand for ordinal preferences; here and in what follows only the order matters. Clearly our Deductive approach does

	$\underline{l}$	$\underline{k}$
$\underline{l}$	$1, 3$	$0, 0$
$\underline{k}$	$0, 0$	$3, 1$

Figure 4.12: Battle of the sexes

not help us to reduce this game in any way as it is entirely symmetric. The game can be represented in its extensive form with imperfect information, as in Figure 4.13. Now though, suppose that Alice has an 'outside option', something that for her falls in between the two possibilities for coordination with Bob. If she chooses to take the option, then play stops, but if she doesn't take the option, then both engage in the existing coordination game. This extended game, with the outside option for Alice, is depicted in Figure 4.14. (Notice that it does not matter what value we put here for X!) The point about this outside option is that, by an argument known as *forward induction*, it enables Alice to ensure that she gets her preferred option without her actually having to *use* the outside option. That is because Bob can reason in the following way: If Alice doesn't take her outside option then, *if she is rational*, it must be that she believes that *I* will opt for my least preferred option in the coordination, i.e. her most preferred option. This sort of reasoning is unfortunately not amenable to the kind of analysis that we proposed as a ground for backward induction, because it involves reasoning about counterfactual situations *that have been eliminated*. If Bob does not choose N, then we would treat this as a public announcement that $\neg N$, and so eliminate it

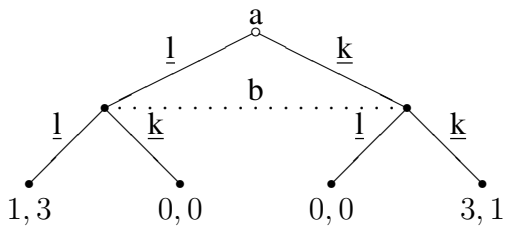


Figure 4.13: Battle of the sexes in extensive form

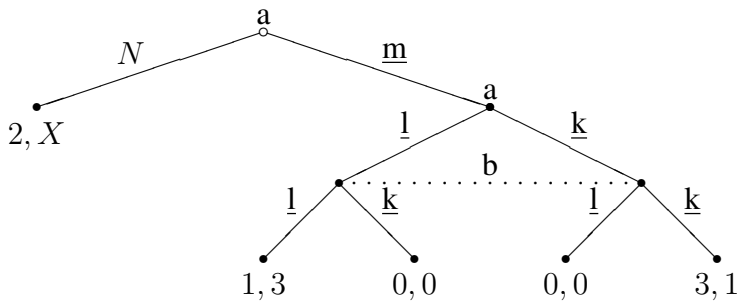


Figure 4.14: Battle of the sexes with an outside option

from all consideration, so that our resulting model would be exactly identical with a model of the smaller game in Figures 4.12 and 4.13. Battle of the sexes with an outside option involves what is known as '*strategic communication*', and this crucially involves remembering the context, *where the game came from*, and so is not directly amenable to the sort of public announcement approach that was so fruitful in the case of games with perfect information. We did extend the idea of modelling moves as public announcements to games with imperfect information, by looking at DEL models for actions. However, that extension also suffers from the same issue of eliminating from consideration nodes (and so outcomes) that are not reached.

We could study instead the strategic form of the game, as depicted in Figure 4.15. Then it can be seen that iterated elimination of weakly dominated strategies mean

	$\underline{l}$	$\underline{k}$
N	$2, X$	$2, X$
$\underline{l}$	$1, 3$	$0, 0$
$\underline{k}$	$0, 0$	$3, 1$

Figure 4.15: Battle of the sexes with an outside option, in normal form

indeed that the only outcome available is $3, 1$, as predicted by the forward induction reasoning (cf. [Osborne and Rubinstein, 1994, p. 110]. However, just as we argued that there is some insight to be gained from studying backward induction on the tree itself, rather than in the strategic form of the game, so we would like to suggest that to better understand strategic communication, from the point of view of dynamic epistemic logic, would mean providing an analysis in terms of the extensive form itself.

We suggest that these two last topics we have looked at, studying solution concepts in terms of lexicographic rationality combined with some belief revision policy, and an analysis of strategic communication within a dynamic epistemic logic framework, are eminently worth of further study.

Summary

In this Chapter we used some of the logical tools from earlier Chapters in order to study *extensive* or '*dynamic*' games. Our principal contribution involved giving conditions, in terms of 'stable belief', for backward induction in generic games of perfect information. This condition can be thought of as a kind of 'belief revision policy': try at all costs to hold on to the belief that players *will* play rationally in the future.

We also considered, for the first time in this work, some solution concepts that are plausible only for a steady-state interpretation of game theory. We showed how to give conditions for such solution concepts in terms of a belief revision policy in concert with lexicographic rationality. We did this by giving a particular example a solution concept, that we called "even-handed trembling-hand equilibrium." Although this is closely

related to trembling-hand equilibrium (otherwise known as 'perfect' equilibrium), of which it is a refinement, it is to our knowledge a new concept. Finally, we illustrated a limitation to the DEL analysis we proposed, by pointing out that it does not yet yield any insight or understanding of the phenomenon of strategic communication.

Summary

"*What does it mean to end anything?*
[long pause]
"*So what is ending?*"
– Jiddu Krishnamurti, 1981

All we permit ourselves in this Summary is a brief recapitulation of each Chapter, followed by a briefer discussion of a set of issues, and questions that have been explicitly left unresolved.

Looking back

The main contribution of **Chapter 1** was to generalise results relating levels of mutual in rationality with numbers of rounds of elimination of non-optimal strategies. There were three directions of generalisation:

1. The first was an extension to *infinite* games, and so to arbitrary (including transfinite) levels of mutual belief in rationality. (We justified this by pointing out that for any arbitrary ordinal α there are games that require α rounds of elimination of not strictly dominated strategies before no more strategies can be eliminated.)

2. Another involved looking at just *how much* logic is required to get some known results. It turned out that the answer is 'not very much', in the sense that we can drop an important axiom about being able to put pieces of information together (saying that if a player believes φ and ψ then she believes their conjunction $\varphi \wedge \psi$) and still get the main result. Similarly, when players are able to put the information together but do not have any kind of introspection concerning their beliefs, the main result still holds.

3. The other generalisation involved considering abstract notions of optimality, so that rather than talk about specific ways of saying when a strategy is 'better' than another, our results hold for all properties that respect a certain property of 'monotonicity'.

In **Chapter 2** we turned our attention to formal languages. We catalogued a variety of languages, and looked at the definability of the key notions of rationality and common belief. We then looked at type-space models, and showed how they connect with the state-space models we had used so far, and which are more familiar from the modal logic literature. The rest of the Chapter was devoted to studying the notion of assumption-completeness. We proved that an infinitary modal language is assumption-complete, and left an open conjecture concerning a language lying between modal and first-order. We take that conjecture to be of independent interest outside of game theory or interactive epistemology, due to the connection between assumption-completeness and Russell's paradox. This technical open question aside, Chapter 2 also left almost untouched conceptual questions concerning the notion of assumption-completeness: most notably, to what extent does it respond to the intuitions expressed behind it, of the 'availability' of a language to a player in a model?

In **Chapter 3** we looked at some aspects of logical dynamics as applied to strategic games. We proposed an interpretation of public announcement actions (cf. [Benthem, 2007b]) as steps in some collective reasoning process. Our aim here was twofold: firstly to look at where models like those used in some of the proofs in Chapter 1 come from; And secondly, to tell some kind of coherent story about this process of common reasoning. Both aims led us to introduce variants of public announcements of rationality. The second also involved introducing, informally, the notions of 'stable equilibrium of beliefs', and to treat the case of *non-monotonic* optimality operators like weak dominance, 'lexicographic rationality'. We showed how 'soft announcements' [Benthem, 2007a] of rationality in this framework give some kind of coherent epistemic analysis of rounds of iteration of eliminating weakly dominated strategies.

The first two directions of generalisation in Chapter 1 both entailed considering *neighbourhood semantics*, in contrast to the relational semantics traditionally used by modal logicians and game-theorists. In Chapter 3, we also generalised some known results about dynamic epistemic logic to the case of neighbourhood models, in particular giving a *reduction axiom* for the basic DEL action modality [Baltag *et al.*, 1999] in a language with *monotonic* modal operators.

Finally in **Chapter 4** we studied some epistemic aspects of *extensive games*. The main contribution there was to give conditions for backwards induction, in terms of *stable belief* and *dynamic* (forward-looking) *rationality*. We looked at how dynamic epistemic logic might be used to analyse extensive-form games with imperfect information. We introduced a solution concept that is a refinement of trembling-hand equilibrium (named 'even-handed'), that we motivated on epistemic grounds. We suggested that both of these solution concepts – backward induction and even-handed trembling hand – could be understood in terms of a belief revision policy combined

with lexicographic rationality. We also showed a limitation in extending the analysis we proposed of backward induction to the interesting problem of an epistemic analysis of strategic communication.

Looking forward

One of the issues raised by our use of neighbourhood semantics is that of the information-processing capacities of players: in neighbourhood models one does not assume that players are able to put information together to the same extent that it is assumed in relational models. We found, in Theorem 1.4, a condition that is just enough to get a certain result about reasoning in the context of a game. Other kinds of interactive reasoning must operate on the basis of other principles which are essentially axioms of some logical system.

This line of thought is in the vein of abstracting away from non-cooperative game theory, which has provided a nice focus for us, but, as we mentioned in the Introduction, does not have a monopoly on interactive reasoning. In that same vein, we left unanswered the question what assumption-completeness means in interactive epistemology generally, irrespective of any game-theoretical application.

And although our account of 'private but common' reasoning (Chapter 3) makes sense in the context of a game, since it just duplicates a given game-theoretical algorithm, if it is to have solid conceptual currency then it should also find some other correlate outside of game theory. This will involve using richer action models from dynamic epistemic logic than the very simple examples we used.

Questions of an increasingly more technical nature that we have raised are:

1. Can dynamic epistemic logic provide an account of strategic communication? (Chapter 4)

2. To what extent can we claim that there is an *equivalence* between

   (a) $1 + \alpha$ rounds of elimination of non-optimal strategies, and

   (b) rationality and α-level mutual belief in rationality?

   (Chapters 1 and 3)

3. Is the modal language with the binder (equivalently: the bounded fragment of first-order logic) assumption-complete? (Chapter 2)

Bibliography

[Apt and Zvesper, 2007] Krzysztof R. Apt and Jonathan A. Zvesper. Common beliefs and public announcements in strategic games with arbitrary strategy sets. Under review. Available from `http://arxiv.org/pdf/0710.3536v2`, 2007.

[Apt, 2007a] Krzysztof R. Apt. Epistemic analysis of strategic games with arbitrary strategy sets. In *Proceedings 11th Conference on Theoretical Aspects of Reasoning about Knowledge (TARK07)*, pages 22–38, 2007. Available from `http://arxiv.org/abs/0706.1001`.

[Apt, 2007b] Krzysztof R. Apt. The many faces of rationalizability. *Berkeley Electronic Journal of Theoretical Economics*, 7(1), 2007. 38 pages.

[Apt, 2007c] Krzysztof R. Apt. Relative strength of strategy elimination procedures. *Economics Bulletin*, 3:1–9, 2007.

[Areces *et al.*, 1999] Carlos Areces, Patrick Blackburn, and Maarten Marx. Hybrid logic is the bounded fragment of first order logic. In R. de Queiroz and W. Carnielli, editors, *Proceedings of 6th Workshop on Logic, Language, Information and Computation, WOLLIC99*, pages 33–50, Rio de Janeiro, Brazil, 1999.

[Aumann and Brandenburger, 1995] Robert J. Aumann and Adam Brandenburger. Epistemic conditions for nash equilibrium. *Econometrica*, 63(5):1161–1180, 1995.

[Aumann, 1974] Robert J. Aumann. Subjectivity and correlation in randomized strategies. *Journal of Mathematical Economics*, 1:67–96, 1974.

[Aumann, 1976] Robert J. Aumann. Agreeing to disagree. *The Annals of Statistics*, 4(6):1236–1239, 1976.

[Aumann, 1985] Robert J. Aumann. What is game theory trying to accomplish? In Kenneth Arrow and S. Honkapohja, editors, *Frontiers of Economics*, pages 28–76. Blackwell, Oxford, 1985.

[Aumann, 1995] Robert J. Aumann. Backward induction and common knowledge of rationality. *Games and Economic Behavior*, 8:6–19, 1995.

[Aumann, 1999] Robert J. Aumann. Interactive epistemology I: Knowledge. *International Journal of Game Theory*, 28:263–300, 1999.

[Aumann, 2006] Robert J. Aumann. War and peace. In Karl Grandin, editor, *The Nobel Prizes 2005*, pages 350–358. Nobel Foundation, Stockholm, 2006.

[Balbiani *et al.*, 2008] Philippe Balbiani, Alexandru Baltag, Hans van Ditmarsch, Andreas Herzig, Tomohiro Hoshi, and Tiago de Lima. 'Knowable' as 'known after an announcement'. *The Review of Symbolic Logic*, 1(3):305–334, October 2008.

[Baltag and Moss, 2004] Alexandru Baltag and Lawrence S. Moss. Logics for epistemic programs. *Synthese*, 139:165–224, 2004.

[Baltag and Smets, 2006] Alexandru Baltag and Sonja Smets. Conditional doxastic models: A qualitative approach to dynamic belief revision. *Electronic Notes in Theorerical Computer Science*, 165:5–21, 2006.

[Baltag and Smets, 2008a] Alexandru Baltag and Sonja Smets. The logic of conditional doxastic actions. In Robert van Rooij and Krzysztof R. Apt, editors, *New Perspectives on Games and Interaction*, volume 4 of *Texts in Logic and Games*, pages 9–31. Amsterdam University Press, 2008.

[Baltag and Smets, 2008b] Alexandru Baltag and Sonja Smets. A qualitative theory of dynamic interactive belief revision. In Giacomo Bonanno, Wiebe van der Hoek, and Michael Wooldridge, editors, *Logic and the Foundations of Game and Decision Theory (LOFT 7)*, volume 3 of *Texts in Logic and Games*, pages 9–58. Amsterdam University Press, 2008.

[Baltag and Smets, 2009] Alexandru Baltag and Sonja Smets. Group belief dynamics under iterated revision: Fixed points and cycles of joint upgrades. In Aviad Heifetz, editor, *Proceedings of the 12th Conference on Theoretical Aspects of Rationality and Knowledge (TARK)*, pages 41–50, 2009.

[Baltag *et al.*, 1999] Alexandru Baltag, Lawrence S. Moss, and Slawomir Solecki. The logic of public announcements, common knowledge and private suspicions. Technical Report SEN-R9922, Centrum voor Wiskunde en Informatica, 1999.

[Baltag *et al.*, 2009] Alexandru Baltag, Sonja Smets, and Jonathan A. Zvesper. Keep 'hoping' for rationality: A solution to the backward induction paradox. *Synthese*, 169(2):301–333, 2009.

[Baltag, 1998] Alexandru Baltag. *A Structural Theory of Sets*. PhD thesis, Indiana University, 1998.

[Barwise, 1988] John Barwise. Three views on common knowledge. In *Proceedings of the second conference on Theoretical aspects of reasoning about knowledge*, pages 365–379, 1988.

[Battigalli and Bonanno, 1999] Pierpaolo Battigalli and Giacomo Bonanno. Recent results on belief, knowledge and the epistemic foundations of game theory. *Research in Economics*, 53:149–225, 1999.

[Battigalli and Siniscalchi, 1999] Pierpaolo Battigalli and Marciano Siniscalchi. Hierarchies of conditional beliefs and interactive epistemology in dynamic games. *Journal of Economic Theory*, 88(1):188–230, September 1999.

[Battigalli and Siniscalchi, 2002] Pierpaolo Battigalli and Marciano Siniscalchi. Strong belief and forward induction reasoning. *Journal of Economic Theory*, 106(2):356–391, October 2002.

[Battigalli, 1997] Pierpaolo Battigalli. On rationalizability in extensive games. *Journal of Economic Theory*, 74(1):40–61, May 1997.

[Benthem and Bezhanishvili, 2007] Johan van Benthem and Guram Bezhanishvili. Modal logics of space. In Marco Aiello, Ian Pratt-Hartmann, and Johan van Benthem, editors, *Handbook of Spatial Logics*, pages 217–298. Springer, 2007.

[Benthem and Liu, 2007] Johan van Benthem and Fenrong Liu. Dynamic logic of preference upgrade. *Journal of Applied Non-Classical Logics*, 17(2):157–182, 2007.

[Benthem and Pacuit, 2006] Johan van Benthem and Eric Pacuit. The tree of knowledge in action: Towards a common perspective. In Guido Governatori, Ian M. Hodkinson, and Yde Venema, editors, *Advances in Modal Logic*, pages 87–106. College Publications, 2006.

[Benthem and Sarenac, 2004] Johan van Benthem and Darko Sarenac. The geometry of knowledge. In *Aspects of Universal Logic, volume 17 of Travaux Log*, pages 1–31, 2004.

[Benthem *et al.*, 2005] Johan van Benthem, Jan van Eijck, and Barteld Kooi. Common knowledge in update logics. In *TARK '05: Proceedings of the 10th conference on Theoretical aspects of rationality and knowledge*, pages 253–261, Singapore, Singapore, 2005. National University of Singapore.

[Benthem *et al.*, 2006] Johan van Benthem, Sieuwert van Otterloo, and Olivier Roy. Preference logic, conditionals, and solution concepts in games. In Henrik Lagerlund, Sten Lindström, and Rysiek Sliwinski, editors, *Modality Matters: Twenty-Five Essays in Honour of Krister Segerberg*, volume 53 of *Uppsala Philosophical Studies*, pages 61–76. Uppsala Universitet, 2006.

[Benthem, 1976] Johan van Benthem. *Modal Correspondence Theory*. PhD thesis, Mathematisch Instituut & Instituut voor Grondslagenonderzoek, Universiteit van Amsterdam, 1976.

[Benthem, 1996] Johan van Benthem. *Exploring Logical Dynamics*. CSLI, Stanford, CA, 1996.

[Benthem, 2001] Johan van Benthem. Games in dynamic epistemic logic. *Bulletin of Economic Research*, 53(4):219–48, October 2001.

[Benthem, 2004] Johan van Benthem. What one may come to know. *Analysis*, 64(2):95–105, 2004.

[Benthem, 2007a] Johan van Benthem. Dynamic logic for belief revision. *Journal of Applied Non-Classical Logics*, 17(2):129–155, 2007.

[Benthem, 2007b] Johan van Benthem. Rational dynamics and epistemic logic in games. *International Game Theory Review*, 9(1):13–45, 2007. (Erratum reprint, 9(2), 377–409).

[Benthem, forthcoming] Johan van Benthem. *Logical Dynamics of Information and Interaction*. Cambridge University Press, forthcoming.

[Bernheim, 1984] B. Douglas Bernheim. Rationalizable strategic behavior. *Econometrica*, 52:1007–1028, 1984.

[Bicchieri, 1989] Cristina Bicchieri. Self-refuting theories of strategic interaction: a paradox of common knowledge. *Erkenntnis*, 30:69–85, 1989.

[Binmore, 1987] Ken Binmore. Modeling rational players, part I. *Economics and Philosophy*, 3:179–214, 1987.

[Binmore, 1996] Ken Binmore. A note on backward induction. *Games and Economic Behavior*, 17(1):135–137, November 1996.

[Blackburn *et al.*, 2001] Patrick Blackburn, Maarten de Rijke, and Yde Venema. *Modal Logic*. Cambridge University Press, Cambridge, UK, 2001.

[Blume *et al.*, 1991] Lawrence Blume, Adam Brandenburger, and Eddie Dekel. Lexicographic probabilities and choice under uncertainty. *Econometrica*, 59(1):61–79, 1991.

[Board, 2002] Oliver Board. Dynamic interactive epistemology. *Games and Economic Behavior*, 49:49–80, 2002.

[Bonanno, 1991] Giacomo Bonanno. The logic of rational play in games of perfect information. *Economics and Philosophy*, 7:37–65, 1991.

[Bonanno, 2002] Giacomo Bonanno. Modal logic and game theory: Two alternative approaches. *Risk, Decision and Policy*, 7(3):309–324, November 2002.

[Brandenburger and Keisler, 2006] Adam Brandenburger and H. Jerome Keisler. An impossibility theorem on beliefs in games. *Studia Logica*, 84(2):211–240, November 2006.

[Brandenburger *et al.*, 2008] Adam Brandenburger, Amanda Friedenberg, and H. Jerome Keisler. Admissibility in games. *Econometrica*, 76(2):307 – 352, 2008.

[Brandenburger, 2003] Adam Brandenburger. On the existence of a "complete" possibility structure. In Nicola Dimitri, Marcello Basili, and Itzhak Giboa, editors, *Cognitive Processes and Economic Behavior*, pages 30–34. Routledge, London, 2003.

[Brandenburger, 2007] Adam Brandenburger. The power of paradox: some recent developments in interactive epistemology. *International Journal of Game Theory*, 35(4):465–492, April 2007.

[Bruin, 2004] Boudewijn de Bruin. *Explaining Games: On the logic of game theoretic explanations*. PhD thesis, ILLC, Amsterdam, 2004.

[Cate, 2005] Balder ten Cate. *Model theory for extended modal languages*. PhD thesis, University of Amsterdam, 2005. ILLC Dissertation Series DS-2005-01.

[Chellas, 1980] Brian F. Chellas. *Modal Logic: An Introduction*. Cambridge University Press, Cambridge, UK, 1980.

[Chen *et al.*, 2007] Yi-Chun Chen, Ngo Van Long, and Xiao Luo. Iterated strict dominance in general games. *Games and Economic Behavior*, 61(2):299–315, 2007.

[Clausing, 2003] Thorsten Clausing. Doxastic conditions for backward induction. *Theory and Decision*, 54:315–336, 2003.

[Davidson, 1980] Donald Davidson. *Essays on Actions and Events*. Clarendon, 1980.

[Dégremont and Roy, 2009] Cédric Dégremont and Olivier Roy. Agreement theorems in dynamic epistemic logic. In Aviad Heifetz, editor, *Proceedings of the 12th Conference on Theoretical Aspects of Rationality and Knowledge (TARK)*, pages 91–98, 2009.

[Dégremont and Zvesper, 2010] Cédric Dégremont and Jonathan A. Zvesper. Dynamics we can believe in. To appear in, 2010.

[Devlin, 1993] Keith Devlin. *The Joy of Sets: Fundamentals of Contemporary Set Theory*. Springer, New York, 1993.

[Ditmarsch *et al.*, 2007] Hans van Ditmarsch, Wiebe van der Hoek, and Barteld Kooi. *Dynamic Epistemic Logic*, volume 337 of *Synthese Library Series*. Springer, New York, 2007.

[Duggan, 2003] John Duggan. A note on backward induction, iterative elimination of weakly dominated strategies, and voting in binary agendas. Manuscript, University of Rochester, 2003.

[Fagin *et al.*, 1995] Ronald Fagin, Joseph Y. Halpern, Moshe Vardi, and Yoram Moses. *Reasoning about knowledge*. MIT Press, Cambridge, MA, 1995.

[Feferman, 1968] Solomon Feferman. Persistent and invariant formulas for outer extensions. *Compositio Mathematica*, 20:29–52, 1968.

[Fine, 1970] Kit Fine. Propositional quantifiers in modal logic. *Theoria*, 36:336–346, 1970.

[Forti and Hinnion, 1989] M. Forti and R. Hinnion. The consistency problem for positive comprehension principles. *Journal of Symbolic Logic*, 54(4):1401–1418, December 1989.

[Friedell, 1969] Morris F. Friedell. On the structure of shared awareness. *Behavioral Science*, 14(1):28–39, 1969.

[Fudenberg and Tirole, 1991] Drew Fudenberg and Jean Tirole. *Game Theory*. MIT, 1991.

[Gerbrandy and Groeneveld, 1997] Jelle Gerbrandy and Willem Groeneveld. Reasoning about information change. *Journal of Logic, Language, and Information*, 6:147–169, 1997.

[Gerbrandy, 1999] Jelle Gerbrandy. *Bisimulations on Planet Kripke*. PhD thesis, ILLC, Amsterdam, 1999.

[Halpern and Lakemeyer, 2001] Joseph Y. Halpern and G Lakemeyer. Multi-agent only knowing. *Journal of Logic and Computation*, 11(1):41–70, 2001.

[Halpern *et al.*, 2007] Joseph Y. Halpern, Dov Samet, and Ella Segev. Defining knowledge in terms of belief: The modal logic perspective. manuscript, 2007.

[Halpern, 2001] Joseph Y. Halpern. Substantive rationality and backward induction. *Games and Economic Behavior*, 37:425–435, 2001.

[Halpern, 2008] Joseph Y. Halpern. A nonstandard characterization of sequential equilibrium, perfect equilibrium, and proper equilibrium. *Int. Journal of Game Theory*, 2008. to appear.

[Hansen *et al.*, 2009] Helle Hvid Hansen, Clemens Kupke, and Eric Pacuit. Neighbourhood structures: Bisimilarity and basic model theory. *Logical Methods in Computer Science*, 5(2):1–38, 2009.

[Hansen, 2003] Helle Hvid Hansen. Monotonic modal logics. Master's thesis, Institute for Logic, Language and Computation, Amsterdam, 2003. ILLC Prepublication Series PP-2003-04.

[Harsanyi, 1968] John C. Harsanyi. Games with incomplete information played by bayesian players, i. *Management Science*, 14:159–182, 1968.

[Heifetz, 1996] Aviad Heifetz. Common belief in monotonic epistemic logic. *Mathematical Social Sciences*, 32:109–123, 1996.

[Heifetz, 1999] Aviad Heifetz. Iterative and fixed point common belief. *Journal of Philosophical Logic*, 28:61–79, 1999.

[Hintikka, 1962] Jaakko Hintikka. *Knowledge and Belief: an introduction to the logic of the two notions*. Cornell University Press, Ithica, NY, 1962.

[Hintikka, 1975] Jaakko Hintikka. Impossible possible worlds vindicated. *Journal of Philosophical Logic*, 4:475–484, 1975.

[Kanger, 1957] Stig Kanger. The morning star paradox. *Theoria*, 23:1–11, 1957.

[Keisler, 2000] H. Jerome Keisler. *Elementary Calculus: An Infinitesimal Approach*. published online, `http://www.math.wisc.edu/~keisler/calc.html`, 2000.

[Kozen, 1983] Dexter Kozen. Results on the propositional mu-calculus. *Theoretical Computer Science*, 27(3):333–354, 1983.

[Kripke, 1959] Saul Kripke. A completeness theorem in modal logic. *Journal of Symbolic Logic*, 24:1–14, 1959.

[Lacan, 1973] Jacques Lacan. *Le Séminaire, Livre XI: Les quatre concepts fondamentaux de la psychanalyse*. Seuil, Paris, 1973. Text compiled by Jacques-Alain Miller.

[Lasseter, 1995] John Lasseter. Toy story, 1995. Pixar Animation Studios.

[Lehrer, 1997] Tom Lehrer. Sociology. available at `http://www.archive.org/details/lehrer`, 1997.

[Levesque, 1990] Hector J. Levesque. All I know: a study in autoepistemic logic. *Artificial Intelligence*, 42(2-3):263–309, 1990.

[Lewis, 1969] David Lewis. *Convention*. Blackwell, Oxford, 1969.

[Lewis, 1973] David Lewis. *Counterfactuals*. Blackwell, Oxford, 1973.

[Lismont and Mongin, 1994] Luc Lismont and Philippe Mongin. On the logic of common belief and common knowledge. *Theory and Decision*, 37:75–106, 1994.

[Lismont, 1994] Luc Lismont. Common knowledge: Relating anti-founded situation semantics to modal logic neighbourhood semantics. *Journal of Logic, Language, and Information*, 3(4):285–302, 1994.

[McKinsey and Tarski, 1944] J.C.C. McKinsey and Alfred Tarski. The algebra of topology. *Annals of Mathematics*, 45:141–91, 1944.

[Monderer and Samet, 1989] Dov Monderer and Dov Samet. Approximating common knowledge with common beliefs. *Games and Economic Behavior*, 1(2), 1989.

[Moore, 1942] G.E. Moore. A reply to my critics. In P.A. Schilpp, editor, *The Philosophy of G.E. Moore*, volume 4 of *The Library of Living Philosophers*, pages 535–677. Northwestern University, Evanston IL, 1942.

[Morgenstern, 1928] Oskar Morgenstern. *Wirtschaftsprognose, Eine Untersuchung ihrer Voraussetzungen und Möglichkeiten*. Springer, 1928.

[Munkres, 1999] James Munkres. *Topology (2nd Edition)*. Prentice Hall, Englewood Cliffs, New Jersey, 1999.

[Nash, 1995] John Nash. Nobel acceptance speech. In Tore Frängsmyr, editor, *The Nobel Prizes*. Nobel Foundation, Stockholm, 1995.

[Neumann and Morgenstern, 1944] John von Neumann and Oskar Morgenstern. *Theory of Games and Economic Behavior*. Princeton University Press, Princeton, NJ, 1944.

[Osborne and Rubinstein, 1994] Martin J. Osborne and Ariel Rubinstein. *A Course in Game Theory*. MIT Press, Cambridge, MA, 1994.

[Pacuit, 2007] Eric Pacuit. Understanding the Brandenburger-Keisler paradox. *Studia Logica*, 86(3), 2007.

[Pearce, 1984] D. G. Pearce. Rationalizable strategic behavior and the problem of perfection. *Econometrica*, 52:1029–1050, 1984.

[Plaza, 1989] Jan A. Plaza. Logics of public communications. In M. L. Emrich, M. S. Pfeifer, M. Hadzikadic, and Z. W. Ras, editors, *Proceedings of the 4th International Symposium on Methodologies for Intelligent Systems*, pages 201–216, 1989.

[Putnam, 1975] Hilary Putnam. The meaning of 'meaning'. *Minnesota Studies in the Philosophy of Science*, 7:131–193, 1975.

[Rabinowicz, 1998] Wlodek Rabinowicz. Grappling with the centipede: defense of backward induction for bi-terminating games. *Philosophy and Economics*, 14:95–126, 1998.

[Reny, 1992] Philip Reny. Rationality in extensive form games. *Journal of Economic Perspectives*, 6:92–100, 1992.

[Rényi, 1955] Alfréd Rényi. On a new axiomatic theory of probability. *Acta Mathematica Academiae Scientiarum Hungaricae*, 6:285–335, 1955.

[Rodriguez, 2007] Raul Leal Rodriguez. Topological update for dynamic epistemic logic. manuscript, 2007.

[Rosenthal, 1981] Robert W. Rosenthal. Games of perfect information, predatory pricing, and the chain store. *Journal of Economic Theory*, 25:92–100, 1981.

[Roy, 2008] Olivier Roy. *Thinking before acting: intentions, logic, rational choice*. PhD thesis, ILLC, Amsterdam, 2008.

[Rushdie, 1981] Salman Rushdie. *Midnight's Children*. Cape, London, 1981.

[Samet, 1990] Dov Samet. Ignoring ignorance and agreeing to disagree. *Journal of Economic Theory*, 52(1):190–207, October 1990.

[Samet, 1996] Dov Samet. Hypothetical knowledge and games with perfect information. *Games and Economic Behavior*, 17:230–251, 1996.

[Samet, 2006] Dov Samet. Agreeing to disagree: The non-probabilistic case. manuscript, 2006.

[Scott, 1970] Dana Scott. Advice on modal logic. In K. Lambert, editor, *Philosophical Problems in Logic*, pages 143–173. Reidel, 1970.

[Segerberg, 1995] Krister Segerberg. Belief revision from the point of view of doxastic logic. *Bulletin of the Interest Group in Pure and Applied Logics*, pages 535–553, 1995.

[Segerberg, 2006] Krister Segerberg. Moore problems in full dynamic doxastic logic. In Jacek Malinowski and Andrzej Pietruszczak, editors, *Essays in Logic and Ontology*, volume 91 of *Poznan Studies in the Philosophy of the Sciences and the Humanities*, pages 95–110. Rodopi, November 2006.

[Selten, 1975] Reinhard Selten. Reexamination of the perfectness concept for equilibrium points in extensive games. *Int. Journal of Game Theory*, 4(1):22–55, 1975.

[Sorensen, 2009] Roy Sorensen. Epistemic paradoxes. In Edward N. Zalta, editor, *Stanford Encyclopedia of Philosophy*. 2009. `http://plato.stanford.edu/archives/spr2009/entries/epistemic-paradoxes/`.

[Stalnaker, 1976] Robert C. Stalnaker. Possible worlds. *Noûs*, 10(1), March 1976.

[Stalnaker, 1994] Robert C. Stalnaker. On the evaluation of solution concepts. *Theory and Decision*, 37:49–73, 1994.

[Stalnaker, 1996] Robert C. Stalnaker. Knowledge, beliefs and counterfactual reasoning in games. *Economics and Philosophy*, 12:133–163, 1996.

[Stalnaker, 1998] Robert C. Stalnaker. Belief revision in games: forward and backward induction. *Mathematical Social Sciences*, 36:31–56, 1998.

[Stalnaker, 2006] Robert C. Stalnaker. On logics of knowledge and belief. *Philosophical Studies*, 128:169–199, 2006.

[Tan and Werlang, 1988] Tommy Chin-Chiu Tan and Sergio Ribeiro da Costa Werlang. The Bayesian foundations of solution concepts of games. *Journal of Economic Theory*, 45(2):370–391, August 1988.

[Tarski, 1955] Alfred Tarski. A lattice-theoretic fixpoint theorem and its applications. *Pacific Journal of Mathematics*, 5:285–309, 1955.

[Walukiewicz, 1995] Igor Walukiewicz. Completeness of Kozen's axiomatization of the propositional μ-calculus. In *Proceedings 10th Annual IEEE Symp. on Logic in Computer Science, LICS'95, San Diego, CA, USA, 26–29 June 1995*, pages 14–24. IEEE Computer Society Press, Los Alamitos, CA, 1995.

[Williamson, 2000] Timothy Williamson. *Knowledge and its Limits*. Oxford University Press, Oxford, 2000.

[Zimper, 2005] Alexander Zimper. Equivalence between best responses and undominated strategies: a generalization from finite to compact strategy sets. *Economics Bulletin*, 3(7):1–6, 2005.

[Zvesper and Pacuit, 2010] Jonathan A. Zvesper and Eric Pacuit. A note on assumption-incompleteness in modal logic. In Giacomo Bonanno, Wiebe van der Hoek, and Benedikt Löwe, editors, *Proceedings of the 8th Conference on Logic and the Foundations of Game and Decision Theory (LOFT08)*. Amsterdam University Press, 2010. To appear.

Samenvatting

De titel van dit proefschrift zegt niet zo veel over de inhoud. We onderzoeken niet precies hoe er met informatie wordt "gespeeld", maar veeleer hoe de informatie die spelers bezitten de uitkomst van het spel bepaalt: hoe ze spelen, met informatie. Zo bekijken we bijvoorbeeld stellingen van de vorm 'als de spelers dit soort informatie hebben, dan maken ze dat soort keuzes'.

De inhoud van dit proefschrift beslaat een serie bijdragen aan de literatuur over *epistemische speltheorie*. We onderzoeken de verbanden tussen geloof en rationele keuze in een interactieve, meerspeler context. Onze nadruk, zoals gebruikelijk onder epistemische speltheoretici, ligt op de zogenaamde *nietcooperatieve* speltheorie, waarin bindende contracten die spelers onderling kunnen aangaan expliciet gemodelleerd worden. We proberen zo veel mogelijk de gehele interactiesituatie in onze spelen te modelleren. We proberen zo veel mogelijk te vermijden aan te nemen dat de spelers exogene informatie hebben over wat andere spelers van plan zijn, bijvoorbeeld gebaseerd op eerdere observaties. We noemen dit de 'one shot' interpretatie. Dit betekend dat het type informatie dat wij beschouwen altijd gaat over de 'rationaliteit' van de spelers, of over informatie over dit soort informatie.

Hoofdstuk 1 dient niet alleen ter introductie van enkele wiskundige modellen die we gebruiken in dit proefschrift, maar bevat ook enkele kleine bijdragen aan een simpele doch fundamentele stelling van de epistemische speltheorie, die de 'hoeveelheid wederzijdse geloof' relateert aan het aantal ronden van interactie van niet optimale strategien. Met wat sociologisch vernis zou je kunnen zeggen dat deze stelling een directe correlatie bevestigt tussen aan de ene kant de mate waarin een groep spelers over 'dezelfde' informatie beschikt, en anderzijds de mate waarin het gedrag van die spelers wordt gecordineerd doordat zij de voorkeuren van de anderen overwegen. Zulk vernis zul je in het proefschrift verder niet aantreffen, het is moeilijk deze zaken bondig in lekentermen samen te vatten. De introductie schetst wel de basale logica van het bewijs van de stelling. Onze eigen kleine bijdragen in hoofdstuk 1 noemen we daar 'generalisaties'. Zo bekijken we hoe de logica uit te breiden is naar het oneindige geval. Dit blijkt ietwat subtiel, en we beargumenteren dat we de zogenaamde 'neighbourhood' (of

de sterk gerelateerde z.g.n. 'topologische') modellen voor geloof moeten gebruiken in plaats van de 'relationele' modellen die gebruikelijk zijn in de epistemische logica. Voor diegenen die bekend zijn met epistemische logica: er zijn hints van een verband met de problemen van logische alwetendheid. In neighbourhood modellen kun je niet zonder meer afleiden uit kennis van φ en het feit dat φ ψ impliceert dat je dan ψ weet.

In hoofdstuk 2 maken we het onderscheid, belangrijk in de logica, tussen syntax (taal) en semantiek (modellen), en bespreken de bijbehorende onderwerpen, zoals *definieerbaarheid*. De technische bijdrage van dat hoofdstuk ligt in het beantwoorden van een fundamentele vraag over het bestaan van een geloofsmodel dat in bepaalde zin 'compleet' is.

In hoofdstuk 3 spelen we met enkele ideen over de *dynamiek* van informatie, en bekijken waar epistemische condities vandaan komen. We introduceren gereedschappen uit de 'dynamisch epistemische logica', en passen die aan voor gebruik in de neighbourhood modellen die we veelvuldig gebruiken in dit proefschrift. We tonen ook het belang aan, voor het begrip van bepaalde speltheoretische voorspellingen, van herzienbaar geloof: dat wil zeggen, het modelleren van situaties waarin een speler iets eerst kan geloven, en later kan leren dat ditgene niet waar is.

In hoofdstuk 4 bekijken we een specifieke speltheoretische situatie waarin de aannamen van de spelers zelf onwaar kunnen zijn in deze zin, en ze verrast kunnen worden door de klaarblijkelijke irrationaliteit van andere spelers. Om deze reden geven we aandacht aan spellen met verscheidene temporele stadia (zogenaamde 'extensive games'). We gebruiken de behandelde gereedschappen uit hoofdstuk 3 om epistemische modellen van zulke situaties te maken, waarin spelers dingen kunnen geloven die later onwaar blijken. (Voor de speltheoreticus: we geven een epistemisch raamwerk in termen van het begrip 'stabiel geloof in dynamische rationaliteit', voor achterwaartse inductie.)

Abstract

"*They can take one small matrix,*
And really do great tricks,
All in the name of sociology"
– Tom Lehrer [1997]

The title of this dissertation is not very informative as to its contents. We do not look exactly at how information is 'played with', but rather at how information the players have affects the play of the game: at how they play, with information. So for example some of the theorems we discuss are of the form, 'if the players have such-and-such information, then they will make such-and-such choices'.

The contents of this dissertation therefore constitute a series of contributions to the literature on *epistemic game theory*. So we study the connections between beliefs and rational choice in an interactive, multi-agent setting. We focus, as has been focussed the attention of epistemic game theorists, only on so-called '*non-cooperative*' game theory, i.e. in which any binding contracts the players can make between themselves must be explicitly modelled in the game. Indeed, as much as is possible we try to let each game be the whole story about the interaction situation, so we generally avoid assuming that players have exogenous information concerning what other players will do, based for example on past observation. We call this the 'one-shot' interpretation. It means that the kind of information we consider is always about the 'rationality' of the players, or information about information of this kind.

In Chapter 1, which also serves to introduce some of the mathematical models that we use in the dissertation, we add a few minor touches to a basic but fundamental theorem in epistemic game theory, which relates the 'level' of 'mutual belief' to the number of rounds of iteration of non-optimal strategies. To put a sociological gloss on that theorem, we could see it as affirming a direct correlation between in the one hand the extent to which a group of players have the 'same' information and in the other the extent to which those players' behaviour is co-ordinated by consideration of

the preferences of others. However, you will not find such gloss on the material in the dissertation, and it's difficult to sum up the issues concisely in non-specialist terms, though the Introduction does briefly sketch part of the basic logic of the argument proving the theorem. The 'minor touches' that form our own contribution in Chapter 1 are called there 'generalisations', and one of those is to look at how the logic extends to the *infinitary* case, where it turns out that there are some subtleties that, we argue, call for so-called 'neighbourhood', or the closely-related 'topological', models for beliefs, rather than the 'relational' models commonly found in epistemic logic. For those familiar with formal epistemology: there are hints of a connection with issues of logical omniscience, as neighbourhood models do not licence the inference that because you know φ and that φ implies ψ, then you know ψ.

In Chapter 2 we make the distinction, important in logic, between syntax (language) and semantics (models), and discuss some issues that arise, like *definability*. The technical contribution of that Chapter is to address a foundational question concerning the existence of belief model that is in a certain sense 'complete'.

In Chapter 3 we play with some ideas about *dynamics* of information, looking at how epistemic conditions might come about. We introduce tools from 'dynamic epistemic logic', that we adapt to the neighbourhood model framework that we often use throughout the dissertation. We also show the importance, for understanding some game-theoretical predictions, of *revisable* beliefs: that is, of modelling situations in which a player might believe something and later learn that it is not true.

In Chapter 4, we look at a particular game-theoretical situation in which the players' assumptions can be violated in this way, in which they can be surprised by the apparent *irrationality* of a player. So we turn our attention to games with distinct temporal stages (so-called 'extensive games'). We use tools explained in Chapter 3 in order to build epistemic models of these situations, in which players can have beliefs and later find out that they are wrong. (For the game-theorist: we provide an epistemic foundation, in terms of a notion of 'stable belief in dynamic rationality', for backward induction.)

Index

Titles in the ILLC Dissertation Series:

ILLC DS-2001-01: **Maria Aloni**
Quantification under Conceptual Covers

ILLC DS-2001-02: **Alexander van den Bosch**
Rationality in Discovery - a study of Logic, Cognition, Computation and Neuropharmacology

ILLC DS-2001-03: **Erik de Haas**
Logics For OO Information Systems: a Semantic Study of Object Orientation from a Categorial Substructural Perspective

ILLC DS-2001-04: **Rosalie Iemhoff**
Provability Logic and Admissible Rules

ILLC DS-2001-05: **Eva Hoogland**
Definability and Interpolation: Model-theoretic investigations

ILLC DS-2001-06: **Ronald de Wolf**
Quantum Computing and Communication Complexity

ILLC DS-2001-07: **Katsumi Sasaki**
Logics and Provability

ILLC DS-2001-08: **Allard Tamminga**
Belief Dynamics. (Epistemo)logical Investigations

ILLC DS-2001-09: **Gwen Kerdiles**
Saying It with Pictures: a Logical Landscape of Conceptual Graphs

ILLC DS-2001-10: **Marc Pauly**
Logic for Social Software

ILLC DS-2002-01: **Nikos Massios**
Decision-Theoretic Robotic Surveillance

ILLC DS-2002-02: **Marco Aiello**
Spatial Reasoning: Theory and Practice

ILLC DS-2002-03: **Yuri Engelhardt**
The Language of Graphics

ILLC DS-2002-04: **Willem Klaas van Dam**
On Quantum Computation Theory

ILLC DS-2002-05: **Rosella Gennari**
Mapping Inferences: Constraint Propagation and Diamond Satisfaction

ILLC DS-2002-06: **Ivar Vermeulen**
A Logical Approach to Competition in Industries

ILLC DS-2003-01: **Barteld Kooi**
Knowledge, chance, and change

ILLC DS-2003-02: **Elisabeth Catherine Brouwer**
Imagining Metaphors: Cognitive Representation in Interpretation and Understanding

ILLC DS-2003-03: **Juan Heguiabehere**
Building Logic Toolboxes

ILLC DS-2003-04: **Christof Monz**
From Document Retrieval to Question Answering

ILLC DS-2004-01: **Hein Philipp Röhrig**
Quantum Query Complexity and Distributed Computing

ILLC DS-2004-02: **Sebastian Brand**
Rule-based Constraint Propagation: Theory and Applications

ILLC DS-2004-03: **Boudewijn de Bruin**
Explaining Games. On the Logic of Game Theoretic Explanations

ILLC DS-2005-01: **Balder David ten Cate**
Model theory for extended modal languages

ILLC DS-2005-02: **Willem-Jan van Hoeve**
Operations Research Techniques in Constraint Programming

ILLC DS-2005-03: **Rosja Mastop**
What can you do? Imperative mood in Semantic Theory

ILLC DS-2005-04: **Anna Pilatova**
A User's Guide to Proper names: Their Pragmatics and Semanics

ILLC DS-2005-05: **Sieuwert van Otterloo**
A Strategic Analysis of Multi-agent Protocols

ILLC DS-2006-01: **Troy Lee**
Kolmogorov complexity and formula size lower bounds

ILLC DS-2006-02: **Nick Bezhanishvili**
Lattices of intermediate and cylindric modal logics

ILLC DS-2006-03: **Clemens Kupke**
Finitary coalgebraic logics

ILLC DS-2006-04: **Robert Špalek**
Quantum Algorithms, Lower Bounds, and Time-Space Tradeoffs

ILLC DS-2006-05: **Aline Honingh**
The Origin and Well-Formedness of Tonal Pitch Structures

ILLC DS-2006-06: **Merlijn Sevenster**
Branches of imperfect information: logic, games, and computation

ILLC DS-2006-07: **Marie Nilsenova**
Rises and Falls. Studies in the Semantics and Pragmatics of Intonation

ILLC DS-2006-08: **Darko Sarenac**
Products of Topological Modal Logics

ILLC DS-2007-01: **Rudi Cilibrasi**
Statistical Inference Through Data Compression

ILLC DS-2007-02: **Neta Spiro**
What contributes to the perception of musical phrases in western classical music?

ILLC DS-2007-03: **Darrin Hindsill**
It's a Process and an Event: Perspectives in Event Semantics

ILLC DS-2007-04: **Katrin Schulz**
Minimal Models in Semantics and Pragmatics: Free Choice, Exhaustivity, and Conditionals

ILLC DS-2007-05: **Yoav Seginer**
Learning Syntactic Structure

ILLC DS-2008-01: **Stephanie Wehner**
Cryptography in a Quantum World

ILLC DS-2008-02: **Fenrong Liu**
Changing for the Better: Preference Dynamics and Agent Diversity

ILLC DS-2008-03: **Olivier Roy**
Thinking before Acting: Intentions, Logic, Rational Choice

ILLC DS-2008-04: **Patrick Girard**
Modal Logic for Belief and Preference Change

ILLC DS-2008-05: **Erik Rietveld**
Unreflective Action: A Philosophical Contribution to Integrative Neuroscience

ILLC DS-2008-06: **Falk Unger**
Noise in Quantum and Classical Computation and Non-locality

ILLC DS-2008-07: **Steven de Rooij**
Minimum Description Length Model Selection: Problems and Extensions

ILLC DS-2008-08: **Fabrice Nauze**
Modality in Typological Perspective

ILLC DS-2008-09: **Floris Roelofsen**
Anaphora Resolved

ILLC DS-2008-10: **Marian Counihan**
Looking for logic in all the wrong places: an investigation of language, literacy and logic in reasoning

ILLC DS-2009-01: **Jakub Szymanik**
Quantifiers in TIME and SPACE. Computational Complexity of Generalized Quantifiers in Natural Language

ILLC DS-2009-02: **Hartmut Fitz**
Neural Syntax

ILLC DS-2009-03: **Brian Thomas Semmes**
A Game for the Borel Functions

ILLC DS-2009-04: **Sara L. Uckelman**
Modalities in Medieval Logic

ILLC DS-2009-05: **Andreas Witzel**
Knowledge and Games: Theory and Implementation

ILLC DS-2009-06: **Chantal Bax**
Subjectivity after Wittgenstein. Wittgenstein's embodied and embedded subject and the debate about the death of man.

ILLC DS-2009-07: **Kata Balogh**
Theme with Variations. A Context-based Analysis of Focus

ILLC DS-2009-08: **Tomohiro Hoshi**
Epistemic Dynamics and Protocol Information

ILLC DS-2009-09: **Olivia Ladinig**
Temporal expectations and their violations

ILLC DS-2009-10: **Tikitu de Jager**
"Now that you mention it, I wonder. . . ": Awareness, Attention, Assumption

ILLC DS-2009-11: **Michael Franke**
Signal to Act: Game Theory in Pragmatics

ILLC DS-2009-12: **Joel Uckelman**
More Than the Sum of Its Parts: Compact Preference Representation Over Combinatorial Domains

ILLC DS-2009-13: **Stefan Bold**
Cardinals as Ultrapowers. A Canonical Measure Analysis under the Axiom of Determinacy.

ILLC DS-2010-01: **Reut Tsarfaty**
Relational-Realizational Parsing