The Temporal Mind

Observations on the logic of belief change in interactive systems

Cédric Dégremont

The Temporal Mind

Observations on the logic of belief change in interactive systems

ILLC Dissertation Series DS-2010-03

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@uva.nl
homepage: http://www.illc.uva.nl/

The Temporal Mind

Observations on the logic of belief change in interactive systems

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 Agnietenkapel
op dinsdag 9 maart 2010, te 12.00 uur

door

Cédric Dégremont

geboren te Clamart, Frankrijk

Promotiecommissie:

Promotor: Prof.dr. J.F.A.K. van Benthem

Overige leden:
Prof.dr. K.R. Apt
Prof.dr. G. Bonanno
Prof.dr. W. van der Hoek
Prof.dr. S. Rahman
Prof.dr. F. Veltman

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

The investigations were supported by a GLoRiClass fellowship of the EU Commission (Research Training Fellowship MEST-CT-2005-020841)

Printed and bound by Ipskamp Drukkers.

ISBN: 978–90–5776–204–8

Contents

Acknowledgments

I'm grateful to:

My supervisor Johan van Benthem for directing me through this process, for the extended feedback on manuscripts and after talks, and the insightful syntheses after our meetings that he would always take the time to write, for teaching me how to write papers and giving me the opportunity to visit Stanford; Krzysztof Apt, Giacomo Bonanno, Wiebe van der Hoek, Shahid Rahman and Frank Veltman for accepting to be on my committee; Giacomo and Wiebe for the detailed and helpful comments on the dissertation; Shahid, for generously according me so much time when I was a master student; Benedikt Löwe for coordinating the Gloriclass project; Karin Gigengack, Tanja Kassenaar, Ingrid van Loon, Peter van Ormondt and Marjan Veldhuisen for the invaluable help with all administrative issues and for finding pleasant places for me to live in; Peter for the precious help with the Dutch abstract; Balder ten Cate, for answering numerous email-questions so fast and with so much patience; Alexandru Baltag, Tomohiro Hoshi, Valentin Goranko, Eric Pacuit and Sonja Smets for how this dissertation could benefit from stimulating exchanges, from precise answers and from useful comments over the years, by email, at Stanford, in Hamburg, Prague or Amsterdam; Ulle Endriss for having taught me a lot, especially on topics that are not visible in this dissertation and that relate to computational social choice; Stéphane Airiau, Amélie Gheerbrant, Sujata Ghosh, Patrick Girard, Umberto Grandi, Davide Grossi (frisbee expert), Daisuke Ikegami, Jens Ulrik Hansen, Tikitu de Jager, Fenrong Liu, Stefan Minica, Marc Staudacher, Joel Uckelman, Fernando Velázquez-Quesada, for valuable discussions and comments — and in some cases working sessions — we had over the years (mostly) in Amsterdam; Tikitu for super-carefully proof-reading the dissertation; Mikaël Cozic, Franz Dietrich, Paul Egré, Conrad Heilmann, Nicolas Maudet for interesting exchanges by email, in Paris, London, Lille or Amsterdam; Guillaume Aucher, Philippe Balbiani, Elise Bonzon, Olivier Gasquet, Emiliano Lorini, François Schwarzentruber, Nicolas Troquard, and especially Andreas Herzig and Jérôme Lang, for

making my two visits to Toulouse at the IRIT so pleasant (and seeing to it that there was something I could eat at lunchtime) and for the interesting discussions we had there and in Amsterdam; Frédéric Jouneau, Laurent Keiff and Sébastien Konieczny for all the interesting sessions of the seminar on belief revision in Lille, for all the thought-provoking exchanges, we had before and after I left Lille; and also Fabienne Blaise, Amandine Briffaut, Laurence Broze, Nicolas Clerbout, Michel Crubellier, Patricia Everaere, Matthieu Fontaine, Emmanuel Genot, Marie-Hélène Gorisse, Emeline Huart, Justine Jacot, Eléonore Le Jallé, Juan Redmond, Caroline Simon and Séverine Vanhoutte for contributing — each in their own way — to make Lille a stimulating environment for the master student I was; Denis Bonnay for the precious comments on this dissertation, particularly those on the introduction, for all the advice, for all the academic conversations and the non-academic ones around lunches, dinners or drinks in Amsterdam or Paris; Olivier Roy for the rich working sessions on the paper behind chapter 4, feedback, help and guidance from day 1 to J-1, with lots of little things that pop up in the process of doing a PhD at the ILLC; my officemate and neighbor Jonathan A. Zvesper for the hardly countable French-speaking academic and non-academic conversations, working, veggie- and vegan- cooking, wine drinking, board game playing sessions we had over these three years — including those sessions about the paper I used ideas from in chapter 6; Lena Kurzen for how fun it has been to work together on the joint paper behind chapter 7 and follow-ups in whichever time zone we were; Jakub Szymanik for the illuminating comments on the paper behind chapter 7; Nina Gierasimczuk for all the great suggestions on this dissertation, all I have learned in our working sessions on the joint paper behind chapter 5 and all the non-academic discussions; Nina and Jakub for reading and giving me feedback, sometimes in the middle of the night, on drafts of this dissertation and all the cheese-fondue-dinners; Andi Witzel for patiently convincing LaTeX magic forces to be reasonable when I could not manage any more; Andi, QiQi and Marta for many pleasant afternoons and evenings; Jarmo Kontinen for his inexhaustible gaming spirit; the core UT crew: Andi, Jakub, Jarmo, Nina, Pietro, QiQi and the core hold'em crew: Jakub, Jarmo and Nina for lots of fun evenings; Alex, Antoine, both Camilles, both Clemes, David, Eva, Joana, Laurent, Lulu, Manu, Marie, Marion, Nicole, Nicolas, Olivia, Salvador, Signe, Sune, Vartan for entertaining moments and rich exchanges; Eva, Katie, Olivia, Vartan for precious words and wisdom; Enzo for being maximally cool in the space of possible cats; Céline for all we shared and for defining 'home' for most of these last five years; my *extended* family and especially my parents for their warm and unconditional support.

Amsterdam — Cédric Dégremont
January, 2010.

Joachim d'Auge se tut et fit la mine de réfléchir.

Le chapelain devina que le duc envisageait de passer à la rébellion ouverte. Le héraut devina la même chose. Le duc devina que les deux autres avaient deviné. Le chapelain devina que le duc avait deviné qu'il avait deviné, mais ne devinait point si le héraut avait lui aussi deviné que le duc avait deviné qu'il avait deviné. Le héraut, de son coté, ne devinait point si le chapelain avait deviné que le duc avait deviné qu'il avait deviné, mais il devinait que le duc avait deviné qu'il avait deviné.

Les Fleurs bleues
Raymond Queneau

Joachim of Auge held his peace and composed his features to look as if he were thinking.

The chaplain guessed that the Duke was considering proceeding to open rebellion. The herald guessed the same thing. The Duke guessed that the other two had guessed. The chaplain guessed that the Duke had guessed that he had guessed, but didn't guess whether the herald had also guessed that the Duke had guessed that he'd guessed. The herald, for his part, couldn't guess whether the chaplain had guessed that the Duke had guessed that he'd guessed, but he did guess that the Duke had guessed that he'd guessed.

The Blue Flowers
Raymond Queneau
(trans. Barbara Wright)

Chapter 1

Introduction

Beethoven changed his mind drastically about the dedication of his third symphony after learning from Ries that Napoleon had declared himself emperor. Where is the logic in that? Well, this is what this dissertation is about — the logic of belief change in interactive systems. What are agents thinking when they interact? And why do they behave in social contexts the way they do? Intelligent interaction is puzzling. Analyzing it requires models that abstract away from the complexity of actual, real-life situations. Economists, computer scientists, and philosophers have long studied important features of rational interaction, and identified crucial concepts to carry out an analysis, usually focusing on a specific dimension of the phenomenon. Decision theory analyses how agents make decisions in situation of uncertainty, given their preferences and their beliefs about the world. Game theory analyses strategic decision making, i.e. how agents make decisions in an interactive environment in which "two or more individuals make decisions that will influence each other's welfare" (Myerson [124]). And closer to philosophy and computer science, belief revision theory analyses how an agent can adjust her beliefs to take into account new information that might be inconsistent with them.

1.1 Logic, belief change and rational agency

This dissertation intends to complement these approaches. We will not deal with the global phenomenon of intelligent agency. Neither will we explore all its dimensions. Rather we will focus on the process of *belief change under new information*, a fundamental part of the reasoning processes at work as intelligent agents interact. A theory of belief change analyses how what agents regard as true about their physical and social environment evolves over time as they receive new unexpected information. In fact belief revision has received attention from a wide range of research fields: multi-agent systems, epistemic game theory and interactive epistemology, (formal) epistemology, philosophy of science and

learning theory. Each field focuses on particular scenarios and introduces models that fits their analyses. This dissertation concentrates on *logical* approaches to belief dynamics. Our aim is two-fold: to develop a logical framework giving a unified logical perspective on the phenomenon and to use this framework to build connections with non-logical approaches concerned with belief change.

Developing logics — often modal logics — to reason about belief change and the reasoning processes it underlies, is a continuation of the epistemic logic program. Just like epistemic logic brings knowledge and belief into the object-language, one can make the notion of doxastic change a first-class citizen in a formal language, giving it a semantics and axiomatizing its principles. In fact, two different families of modal logics of belief change exist: doxastic temporal logics and dynamic doxastic logics. They offer different but complementary perspectives. Doxastic temporal logics give a global view of all the possible evolutions of an agent's belief, while dynamic doxastic logics describe local updates that transform a doxastic model into a new one, modeling informational signals as generic 'event models'. To bring a unified logical perspective on belief change, this dissertation will work precisely at their interface.

Our first step is to systematically compare the two logics both at a structural and at a syntactic level. Chapter 2 gives representation theorems linking the dynamic and the temporal approaches, shedding light on the assumptions about agents behind dynamic doxastic logics, while Chapter 3 has a complete logic for a system that merges temporal and dynamic logics of beliefs. From this logical viewpoint we can then build connections with other research fields.

Our three connections go toward interactive epistemology, learning theory and strategic reasoning in games. Interactive epistemology, which constitutes a foundational layer for (epistemic) game theory, deals with interactive or higher-order reasoning: how agents reason about what other agents believe, and reason about what other agents believe about what other agents believe, etc... Agreement results and their dynamic companions are among the core results in interactive epistemology. They study the conditions under which agents can 'agree to disagree' and whether — if they disagree — communication will solve the disagreement. Chapter 4 gives a logical look at these issues, proving static and dynamic agreement results for qualitative structures, and providing syntactic counterparts in formal proofs. Next, formal learning theory offers a mathematical perspective on the epistemological basis of inductive reasoning. Important questions include conditions under which an agent can reliably converge to some correct conjecture about its environment. To make a connection, Chapter 5 gives a logical perspective on the important special case where the possible languages are sets of natural numbers. It shows that checking if a class of languages can be learned inductively is really checking if some formula of an appropriate modal language holds in a certain doxastic-temporal model. Chapter 6 then deepens the analysis of agency over time, introducing logical languages for reasoning about knowledge and belief change in extensive games of imperfect information, and about simple types

of strategic reasoning. Finally Chapter 7 broadens the framework developed in this way in two natural directions: agents' preferences, and coalitional power for groups of agents. This links our analysis to two further fields: cooperative game theory, and social choice theory.

Throughout the dissertation we are concerned with different possible languages that can describe important qualitative structures. Indeed, even if we do fix a class of structures, such logical design questions arise. These choices of languages are made based on external feasibility constraints (decidability, computational complexity, conciseness of the language), but also on the notions and type of reasoning one would like to capture. To check whether some modal languages satisfy all these criteria, questions of definability, expressive power and completeness are at stake, and we will be concerned with them throughout, especially when linking up with other frameworks. Along the way, another recurrent theme is computational complexity, feasibility, and the cases where infeasibility strikes. These questions occur from Chapter 3 onwards. Chapters 3 to 6 study languages for reasoning about belief change, our central topic. But we also perform such analyses for logics for coalitional power and preferences in Chapter 7.

In the next section we discuss briefly what intelligent interaction is about in general, to give the broader context for this dissertation. We discuss the conceptual assumptions behind our logical modeling, and briefly contrast the latter with other approaches: (informal in natural language), and quantitative (probabilistic).

1.2 Perspectives on intelligent interaction

Intelligent agents can be taken to include both human and artificial agents. As agents they are not pure passive observers of their environment, they can act on it and change it. By 'intelligent' we simply mean that they form representations of their environment (they entertain beliefs about it), and that they will revise these representations as they receive and process new information or as they reason about their environment (make assumptions, inferences).

We often have in mind interactive systems in which agents' actions affect a common environment and/or the beliefs other agents have about it. Moreover, we sometimes focus on strategic interaction, that considers rational agents that have preferences about the state of the environment and act according to them, pursuing objectives of their own; and on coalitional (or cooperative) interaction, in which certain actions can only be performed by a group of agents rather than an individual agent.

Here are some concrete intuitions and illustrations for some of the foundational notions we will encounter.

One foundational notion is that of a state of the environment. It might be

taken to include both facts that are external to the agents (the weather, the time, the coordinates of some monument, the distribution of hole cards in texas hold'em...) and internal (their preferences, their capacities...). In the simplest case agents simply entertain beliefs about the state of the environment. For one agent the beliefs of the other agents are also part of the environment and she might entertain beliefs about them. Still, we will sometimes use the term 'first-order belief' to refer to beliefs about the environment itself (rather than beliefs about other agents' beliefs) and use the notion of higher-order belief to refer to beliefs about other agents' beliefs.

For sufficiently well controlled systems, one can distinguish between information (or knowledge) and beliefs. In an online poker game, the amount of chips in front of the different players or the cards in the deck can be said to be 'solid' (or 'hard') information: agents might be said to know exactly how much money each agent has left in her stack. Similarly in a diagnostic process the body temperature or the heart rate of a patient may be considered solid information. On the other hand, an agent might entertain beliefs about the cards that her opponent is holding — maybe because she finds it unlikely that her opponent would play certain hands the way he did so far, or because of past observations about the way this player bets etc. But unless she has actually seen her opponent's hole cards, these are simply beliefs rather than solid information. Similarly a particular diagnosis might be the most natural way of explaining some given symptoms, but at least as long as other (maybe less likely) diagnoses would also explain the symptoms, they cannot be ruled out definitely. In this sense a diagnosis is often a belief, rather than solid information. Our models of agents' beliefs take this distinction into account.

But modeling static beliefs (and knowledge) is only enough to describe the state of mind of an agent at a precise point in time. To account for agents that might change their mind as they learn new information and reason about the interactive system, and the corresponding long-term evolution of their beliefs, informational dynamics themselves need to be analyzed. These include noisy observation, communication, public announcement, inductive inference, strategic reasoning etc. We mentioned that for each phenomenon specific formal frameworks were considered — we will get into their details later — and that this dissertation was trying to contribute to the development of a unifying *logic-based* framework in which all such scenarios can be analyzed.

Even so, the choice of a logic-based approach might raise two questions: why not an informal approach couched in natural language? Or why not go the way of science, and take a probabilistic approach? These are important questions, but in this introduction we can only offer very brief answers:

Logic and the role of conceptual analysis. In the philosophical literature, models of rational interaction are frequently specified informally and the corre-

sponding analysis is also commonly carried out in natural language. Examples are Lewis [117] about assigning knowledge to agents, and Rawls [137] about the decision-making of agents in the "original position", i.e. in situations of incomplete information. They are sometimes presented in a mixture of natural and formal language with Levi [116], on the truth-value of subjunctive conditionals, as an example.

By choosing to rather give formal definitions of our models and develop a logical analysis, we do not escape the need for conceptual analysis. To start with, abstracting from concrete situations, to fix more abstract models, is in itself a conceptual step whose correctness or legitimacy can only be checked or argued for by informal means, even if one were to use mathematical or experimental results to back up one's analysis. Moreover by carrying out a logical analysis of a phenomenon such as belief change and of the reasoning processes in which it plays an important role, we pursue their conceptual analysis.

As for the benefits of this extra effort, we think that if agreement is reached on some logical model as an abstract representation of a particular class of phenomena, some conceptual questions can be given a logical formulation, and informal argumentation can be replaced by mathematical proofs in well-controlled systems. Thus one can give unambiguous definitions and prove rigorous conclusions.

Logic and quantitative approaches. To put it roughly (Appendix A has more details) quantitative approaches work with real numbers: beliefs are encoded as probability spaces (and possibly information partitions) and preferences by utility functions, while qualitative approaches can work with as little as relations. In our logics, beliefs are usually encoded by indistinguishability relations (equivalently, information partitions) and/or plausibility orderings, while preferences are encoded by total pre-orders. Probabilistic approaches use much richer models than qualitative ones, assuming that agents are able to elicit probability functions, e.g. by ranking whole (probabilistic) lotteries. This allows for more fined-grained decision-making, but it may put unrealistic demands on agents. In our qualitative approach agents should simply able to say which of two situations (states of the environment) they think is more likely, and we will even allow the case where states of our models remain incomparable. An important question for us is then, how much of the insight about intelligent agency that quantitative, probabilistic approaches gave us remain in a qualitative, logical setting. A concrete encounter between logical and quantitative approaches will take place in chapter 4 on agreement theorems, where we re-analyze the classic results of Aumann [13] in dynamic logics of belief.

We hope chiefly that this section gave the reader some intuition about our use of basic concepts such as 'knowledge' and 'belief' and about the general context (intelligent interaction) in which this dissertation works. We will now present the logical systems for knowledge and belief that constitute the foundations of the

framework, to be developed in Chapters 2 and 3, for reasoning about their evolution as informational processes unfold. We will meet the epistemic-plausibility models that play a crucial role throughout the dissertation, and a typical doxastic language to reason about them, giving on the way some background about their two historical roots.

1.3 Logics of knowledge and belief

The aim of this section is to familiarize the reader with epistemic-plausibility models (Baltag and Smets [16]). They constitute a cornerstone of our whole analysis of belief change, since the meaning we assign to the notion of belief will be given with respect to these models. We try to give a precise picture of epistemic and doxastic logics that can be interpreted on them: languages and their semantics, and axiomatic proof-systems. (Models and languages to reason about notions of *evolution* or *dynamics* of belief will be given in Section 1.6). This modal logic approach based on epistemic plausibility models has its roots in two independent traditions. It inherits the semantic idea of plausibility orderings to represent the conditional beliefs of an agent from classical AGM [3]-type approaches to belief revision and more precisely from Grove's system of spheres (Subsection 1.3.1). It inherits the idea of and the methods for internalizing the notions of knowledge and belief into the object language from the epistemic logic program of Hintikka and its multi-agent continuations (Subsection 1.3.2). After giving some details on both traditions, we finally present the epistemic plausibility models and conditional doxastic logics recently developed by building on the preceding ideas (Subsection 1.3.3).

1.3.1 AGM and plausibility orderings

The classical formal theories of belief change starting with Alchourrón et al. [3] represent the beliefs of a particular agent as a set of propositional formulas (belief set). Subsequent literature has considered the possibility of distinguishing between basic and derived beliefs by introducing belief base revision [96]. A belief base is a finite set of formulas such that its deductive closure is a belief set. But earlier approaches assumed the belief set to be closed under some well-behaved operation of logical consequence (such as the classical consequence relation).

In the context of belief sets an operation of revision maps pairs composed of a closed set of propositional formulas (the initial beliefs of the agent) and of a propositional formula (the incoming information) to a new closed set of propositional formulas (the new beliefs after revision). AGM [3] introduced a set of postulates that should characterize a reasonable operation of revision. The postulates require among other things that the new information should be accepted by the agent (success), that the new belief set should be consistent provided the

incoming information is consistent (consistency) and that old beliefs should be changed in a minimal way. For a complete list of the AGM postulates, the reader is referred to [79] or [3].

Different representation theorems have been proposed in the literature. One of them has been of particular importance for subsequent developments in modal logics of belief change. Grove [90], building on Lewis [118] has a representation result for the AGM postulates in terms of *systems of spheres*. For the sake of clarity and to avoid any later ambiguity between two slightly different ways of modeling similar notions, in this section, we refrain from introducing any explicit formalism and present results informally. Let a state of the world be (or come with) a complete specification of the state of environment, i.e. about all non-doxastic facts. A sphere is a set of states of the world. Given a propositional formula A, let $||A||$ be the set of states at which A holds. A system of spheres centered on a set of states X is a collection of spheres such that:

1. every two sets are comparable with respect to inclusion;
2. X is a minimal element of this collection with respect to inclusion;
3. the collection contains the set of all possible states of the world, and
4. for every sentence A of the propositional language, there is a smallest sphere S_A intersecting $||A||$.

Grove [90] proved the following result:

Theorem 1.1 (Grove [90])**.** *The following are equivalent*

1. *$\star$ is a revision operation satisfying the AGM postulates.*
2. *For each belief set K, there is a sphere system centered on the set of states satisfying K, such that for all propositional formulas A, $\varphi \in K \star A$ iff φ holds in all states in the intersection of $S_A \cap ||A||$.*

As a conclusion, it is interesting to note that the preceding representation is equivalent to having a total pre-order on states of the world such that every (propositionally definable) non-empty subset has minimal elements. This way of encoding beliefs and conditional beliefs is indeed at the root of models of belief and knowledge introduced in the context of modal logics, that we will take as the basis of our analysis and which make crucial use of such plausibility pre-orders.

1.3.2 Epistemic models and epistemic logics

Twenty years before the seminal paper of Alchourrón, Gärdenfors and Makinson [3] on theory change, Hintikka [102] carried out an important philosophical project

and by doing so introduced formal models that are one of the roots of the tradition this dissertation is in line with. His idea was to import epistemic and doxastic notions into the object language. The aim was thus to develop a logic of such epistemic attitudes giving a formal foundation to "criteria of consistency for [...] sets of [epistemic] statements" or, equivalently, formal foundations to a notion of consequence between epistemic statements.

Epistemic models, as introduced by Hintikka [102], compactly represent the information the agents have about the world (what they know, or 'first-order information'), and about the information possessed by the other agents (what they know about other agents' information, or 'higher-order information'). In what follows, $N = \{1, \ldots, n\}$ is a fixed finite set of agents.

Definition 1.2 (Epistemic Models)**.** *An* epistemic model $\mathcal{M}$ *based on a set of agents* N *is of the form* $(W, (\sim_i)_{i \in N}, V)$*, where* $W \neq \emptyset$*, for each* $i \in N$*,* $\sim_i$ *is a binary equivalence relation on* W*, and* $V : \textsc{prop} \to \wp(W)$*.*

We write $|\mathcal{M}| = W$ to refer to the domain of model $\mathcal{M}$. We refer to a pair $(\mathcal{M}, w)$ with $w \in |\mathcal{M}|$ as a pointed model. An epistemic plausibility *frame* $\mathcal{F}$ is an epistemic plausibility model with the valuation V omitted.

Intuitively $\sim_i$ encodes i's uncertainty: if $s \sim_i t$, then if the actual world were s then i would consider it possible that the world is actually t. Finally we note that we often write $\mathcal{K}_i[w] := \{v \in W \mid w \sim_i v\}$ to denote i's information cell at w. The basic epistemic language is defined as follows:

Definition 1.3 (Epistemic Language)**.** *The epistemic language* $\mathcal{L}_{EL}$ *is defined as follows:*

$$\varphi ::= p \mid \neg\varphi \mid \varphi \wedge \varphi \mid K_i\varphi,$$

where i *ranges over* N*, and* p *over a countable set of proposition letters* PROP*.*

The propositional fragment of this language is standard, and we write $\bot$ for $p \wedge \neg p$ and $\top$ for $\neg\bot$. A formula $K_i\varphi$ should be read as "i knows that φ". We write $\langle i \rangle \varphi$ or $\hat{K}_i\varphi$ for $\neg K_i \neg \varphi$. $\mathcal{L}_{EL}$ is interpreted on epistemic models as follows.

Definition 1.4 (Truth definition)**.**

$$\begin{array}{lll}
\mathcal{M}, w \Vdash p & \text{iff} & w \in V(p) \\
\mathcal{M}, w \Vdash \neg\varphi & \text{iff} & \mathcal{M}, w \nVdash \varphi \\
\mathcal{M}, w \Vdash \varphi \wedge \psi & \text{iff} & \mathcal{M}, w \Vdash \varphi \text{ and } \mathcal{M}, w \Vdash \psi \\
\mathcal{M}, w \Vdash K_i\varphi & \text{iff} & \text{for all } v \text{ such that } w \sim_i v \text{ we have } \mathcal{M}, v \Vdash \varphi
\end{array}$$

The important definition is that of knowledge K_i: a formula is *known* to i if the formula is true in all states that i considers possible. [72, ch.2] is a detailed introduction.

Axiomatization. The set of formulas of $\mathcal{L}_{EL}$ valid over the class of all epistemic models can be axiomatized as follows:

PL	$\vdash \varphi$, for all classical propositional tautologies φ
Nec	If $\vdash \varphi$, then $\vdash K_i\varphi$
$(\mathbf{K}K_i)$	$\vdash K_i(\varphi \to \psi) \to (K_i\varphi \to K_i\psi)$
$(\mathbf{T}K_i)$	$\vdash K_i\varphi \to \varphi$
$(\mathbf{4}K_i)$	$\vdash K_i\varphi \to K_iK_i\varphi$
$(\mathbf{5}K_i)$	$\vdash \langle i\rangle\varphi \to K_i\langle i\rangle\varphi$
MP	If $\vdash \varphi \to \psi$ and $\vdash \varphi$ then $\vdash \psi$

Table 1.1: Axiom system **EL**.

$(\mathbf{4}K_i)$ and $(\mathbf{5}K_i)$ are referred to as positive and negative introspection respectively. On relational structures they characterize transitive and Euclidean uncertainty relations.

The following is a basic result in modal logic. Blackburn et al. [39], ch.4 has historical and formal details.

Theorem 1.5 (see e.g. Blackburn et al. [39])**. EL** *is strongly complete with respect to the class of epistemic models.*

Richer languages include modalities for group epistemic notions such as common knowledge. Here is an example of such a language considered in [72]:

Definition 1.6 (Multi-Agent Epistemic Language)**.** *The multi-agent epistemic language* $\mathcal{L}_{MEL}$ *is defined as follows:*

$$\varphi ::= p \mid \neg\varphi \mid \varphi \wedge \varphi \mid K_i\varphi \mid E_G\varphi \mid D_G \mid C_G\varphi,$$

where i *ranges over* N*,* p *over a countable set of proposition letters* PROP *and* $\emptyset \neq G \subseteq N$*.*

A formula $E_G\varphi$ is read as "each agent in group G knows that φ", $D_G\varphi$ as "it is distributed knowledge among group G that φ", and $C_G\varphi$ as "it is common knowledge among group G that φ". To interpret this language we use the following notion:

Definition 1.7. *For each* $G \subseteq I$*, let* $\sim_G^*$ *be the reflexive-transitive closure of* $\bigcup_{i\in G} \sim_i$*. Let* $[w]_G^* = \{w' \in W \mid w \sim_G^* w'\}$*.*

These formulas are interpreted in epistemic models as follows:

Definition 1.8 (Truth definition)**.**

$$\begin{array}{lll} \mathcal{M}, w \Vdash E_G\varphi & \text{iff} & \forall v\ \forall i \in G\ (\text{if } w \sim_i v \text{ then } \mathcal{M}, v \Vdash \varphi) \\ \mathcal{M}, w \Vdash D_G\varphi & \text{iff} & \forall v\ (\text{if } (w,v) \in \bigcap_{i\in G} \sim_i \text{ then } \mathcal{M}, v \Vdash \varphi) \\ \mathcal{M}, w \Vdash C_G\varphi & \text{iff} & \forall v\ (\text{if } w \sim_G^* v \text{ then } \mathcal{M}, v \Vdash \varphi) \end{array}$$

Axiomatization. The set of formulas of $\mathcal{L}_{MEL}$ valid over the class of all epistemic models can be axiomatized as follows:

$\mathbf{E_G}$	$\vdash E_G\varphi \leftrightarrow \bigwedge_{i\in G} K_i\varphi$
$\mathbf{D_{\{i\}}}$	$\vdash D_{\{i\}}\varphi \rightarrow K_i\varphi$
$\mathbf{K_i/D_G}$	$\vdash (\bigvee_{i\in G} K_i\varphi) \rightarrow D_G\varphi$
$\mathbf{NecD_G}$	If $\vdash \varphi$, then $\vdash D_G\varphi$
$(\mathbf{K}D_G)$	$\vdash D_G(\varphi \rightarrow \psi) \rightarrow (D_G\varphi \rightarrow D_G\psi)$
$(\mathbf{T}D_G)$	$\vdash D_G\varphi \rightarrow \varphi$
$(\mathbf{4}D_G)$	$\vdash D_G\varphi \rightarrow D_GD_G\varphi$
$(\mathbf{5}D_G)$	$\vdash \neg D_G\varphi \rightarrow D_G\neg D_G\varphi$
$\mathbf{C_GFP}$	$\vdash C_G\varphi \rightarrow E_G(\varphi \wedge C_G\varphi)$
$\mathbf{C_GIR}$	If $\vdash \varphi \rightarrow E_G(\varphi \wedge \psi)$ then $\vdash \varphi \rightarrow C_G\psi$

Table 1.2: Axiom system **MEL**.

The following result has several sources. For distributed knowledge see [71, 104]. For common knowledge see Kozen and Parikh [110]'s completeness proof for PDL. Fagin et al. [72], ch.3 has details for both common knowledge and distributed knowledge.

Theorem 1.9 (see e.g. Fagin et al. [72])**. MEL** *is weakly complete with respect to the class of epistemic models.*

1.3.3 Epistemic plausibility models and doxastic logics

Drawing on both the modal, relational approach to knowledge described previously and the semantic models developed in the context of AGM [3] style belief revision theory (such as Grove [90] spheres), Baltag and Smets [16], van Benthem [29], Board [40] and van Ditmarsch [66] have developed new relational models on which both beliefs and conditional beliefs could be interpreted. In this dissertation we will be specifically interested in *epistemic plausibility models* as introduced by Baltag and Smets [16].

In such models agents' knowledge (information) will still be encoded by a collection of uncertainty relations $\sim_i$. But they will also carry a collection of pre-orders $\leq_i$ between worlds standing for plausibility relations that encode the current prior (conditional) beliefs of the agents.

Definition 1.10 (Epistemic Plausiblity Model [16])**.** *An* epistemic plausibility model $\mathcal{M} = \langle W, (\leq_i)_{i\in N}, (\sim_i)_{i\in N}, V\rangle$ *has* $W \neq \emptyset$*, for each* $i \in N$*,* $\leq_i$ *is a pre-order on* W *and* $\sim_i$ *is a binary equivalence relation on* W*, and* $V : \text{PROP} \rightarrow \wp(W)$*.*

The relation $w \leq_i w'$ means that w is considered at least as plausible as w' by agent i. Intuitively, the plausibility pre-orders encode the prior beliefs of agents.

We will often consider plausibility relations to be total, but when we think of beliefs in terms of *multi-criteria decisions*, a pre-order allowing for incomparable situations may be all we get [70]. We will thus sometimes state our results for both total and arbitrary pre-orders. We write $a \simeq b$ ('indifference') if $a \leq b$ and $b \leq a$, and $a < b$ if $a \leq b$ and $b \not\leq a$.

Beliefs for i at w are interpreted as truth in the minimal states of i's information partition at w, in other words a belief operator for i will then be necessity with respect to the most plausible states (i.e. the $\leq_i$-minimal elements) of i's information partition. To guarantee that such minimal elements always exist, we will assume that the epistemic plausibility models satisify *local well-foundedness* or a stronger constraint *well-foundedness*.

Definition 1.11 (Local well-foundedness)**.** *A plausibility pre-order satisfies:*

- **Local well-foundedness**. *If for all $w \in W$, all $i \in N$, and for all X such that $\emptyset \subset X \subseteq \mathcal{K}_i[w]$, X has $\leq_i$-minimal elements.*
- **Well-foundedness**. *If for all X such that $\emptyset \subset X \subseteq W$ and all $i \in N$, X has $\leq_i$-minimal elements.*

$\mathcal{M}$ *satisfies* (Local) Well-foundedness *if every plausibility pre-order has the corresponding property.*

We introduce a few useful shortcuts before we turn to the languages and their truth conditions.

Definition 1.12 ((A priori/a posteriori) Most plausible elements)**.**

- *For all $X \subseteq W$, let $\beta_i(X) = \min_{\leq_i}(X) = \{w : w \text{ is } \leq_i\text{-minimal in } X\}$.*
- *For all $w \in W$, let $\mathcal{B}_i[w] = \beta_i(\mathcal{K}_i[w])$.*

We write $w \rhd_i^{\mathcal{B}} v$ iff $v \in \mathcal{B}_i[w]$, and $w \rightarrow_i^X v$ iff $v \in \beta_i(\mathcal{K}_i[w] \cap X)$.

Intuitively $\beta_i(X)$ are the *a priori* most plausible elements of a set, ignoring the information partitions. $\mathcal{B}_i[w]$ gives the states i considers most plausible, conditional on the information he possesses at w, i.e. conditional on $\mathcal{K}_i[w]$. The relation $w \rightarrow_i^X v$ maps w to all states i considers most plausible, conditional on the information he possesses at w *and* on a given subset X.

Epistemic plausibility models are now ready to support a natural doxastic-epistemic language:

Definition 1.13 (Basic doxastic-epistemic language)**.** *The language $\mathcal{L}_{DOX}$ is defined as follows:*

$$\varphi ::= p \mid \neg\varphi \mid \varphi \wedge \varphi \mid K_i\varphi \mid B_i^{\varphi}\varphi,$$

where i ranges over N, and p over a countable set of proposition letters PROP.

The formula $B_i^{\varphi}\psi$, should be read "conditionally on φ, i believes that ψ." These formulas are interpreted in epistemic plausibility models as follows:

Definition 1.14 (Truth definition)**.** *We write $||\varphi||^{\mathcal{M}}$ for $\{w \in |\mathcal{M}| \ : \ \mathcal{M}, w \Vdash \varphi\}$. We omit $\mathcal{M}$ when it is clear from the context.*

$$
\begin{array}{lll}
\mathcal{M}, w \Vdash p & \text{iff} & w \in V(p) \\
\mathcal{M}, w \Vdash \neg\varphi & \text{iff} & \mathcal{M}, w \nVdash \varphi \\
\mathcal{M}, w \Vdash \varphi \wedge \psi & \text{iff} & \mathcal{M}, w \Vdash \varphi \text{ and } \mathcal{M}, w \Vdash \psi \\
\mathcal{M}, w \Vdash K_i\varphi & \text{iff} & \text{for all } v \text{ such that } w \sim_i v \text{ we have } \mathcal{M}, v \Vdash \varphi \\
\mathcal{M}, w \Vdash B_i^{\psi}\varphi & \text{iff} & \text{for all } v \text{ such that } w \rightarrow_i^{||\psi||^{\mathcal{M}}} v \text{ we have } \mathcal{M}, v \Vdash \varphi
\end{array}
$$

Simple belief conditional only on i's information at a state w can be defined using the conditional belief operator: $B_i\varphi = B_i^{\top}\varphi$, since:

$$\mathcal{M}, w \Vdash B_i^{\top}\varphi \text{ iff } \forall v \ (\text{if } w \rhd_i^{\mathcal{B}} v \text{ then } \mathcal{M}, v \Vdash \varphi).$$

For details about the axiomatization of the doxastic logic $\mathcal{L}_{DOX}$ and similar ones, the reader is referred to [40, 17].

Let us now give some intuition about the use of this language using Example 1.15 which represents a simple doxastic-epistemic situtation that we will put to work in the next chapter. Here is how to read Figure 1.1.

Reading the figures. In Figure 1.1 (and the one involved in the continuation of this example in Chapter 2), the actual state is the shaded one. Epistemic equivalence classes are represented by rectangles or ellipses. We use $<$ to display the strict plausibility ordering within such classes. Our example assumes that all agents have the same plausibility ordering. The agent i believes φ at w is interpreted as φ holds in the i-most plausible states within i-information partition $K_i[w]$. An agent's beliefs at the actual state are thus displayed by an arrow from the actual state to the ones she considers most plausible, often just one. Thus, an arrow from x to y labelled by the agent *Denis* means that y is the $\leq_e$-minimal state within $K_e[x]$. Finally, we omit reflexive arrows throughout.

Example 1.15. ***Knowing about the Wii party.*** *Céline and Enzo would like to invite Denis to their Wii party. The party has been decided but none of them has informed Denis yet. Denis considers it a priori more plausible that no Wii party is taking place unless informed otherwise. This initial situation is common knowledge between Céline and Enzo. In the following figures, plain rectangles (or ellipses) will represent Denis' epistemic partition, dashed ones Enzo's and dotted ones Céline's; w and $\overline{w}$ are state names.*

We can check e.g. that $\mathcal{M}, w \Vdash \neg(K_d p \vee K_d \neg p)$ (Denis does not know whether a Wii party is planned), $\mathcal{M}, w \Vdash B_e K_c \neg(K_d p \vee K_d \neg p)$ (Enzo believes that Céline knows that Denis does not know whether a Wii party is planned), $\mathcal{M}, w \Vdash p \wedge B_d \neg p$ (Denis wrongly believes that no Wii party is planned) and $\mathcal{M}, w \Vdash B_d^p K_e p$ (Denis believes that if there is a Wii party then Enzo knows it).

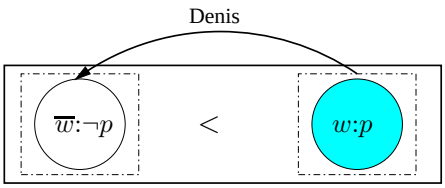

Figure 1.1: No Wii Party unless stated otherwise. Initial model.

We use the notion of *doxastic plausibility models* to refer to epistemic-free epistemic plausibility models. Formally:

Definition 1.16 (Doxastic Plausibility Model)**.** *A* doxastic plausibility model $\mathcal{M} = \langle W, (\preceq_i)_{i\in N}, V\rangle$ *has* $W \neq \emptyset$*, while, for each* $i \in N$*,* $\preceq_i$ *is a pre-order on* W*, and* $V : \text{PROP} \to \wp(W)$*.*

These models are useful to reason on the model-theoretic level when we would like to focus on the plausibility orderings and their dynamics. Let us remark that, in the literature [40, 29], epistemic-free models are considered in which the plausibility ordering is taken to depend on the states. Without any additional assumptions this approach is very general and allows for very diverse types of agents, e.g. non-introspective agents.

A last approach considered in the literature consists in unifying the epistemic relation and the plausibility relation in a single local plausibility ordering $\trianglelefteq$. The intuition is that the states which are epistemically possible at state w are the ones that are $\trianglelefteq$-comparable with w, while i believes φ at w is interpreted as φ holding in the $\trianglelefteq$-minimal elements of $\{v \mid v \trianglelefteq w \text{ or } w \trianglelefteq v\}$. We refer to these models as 'unified' or 'local' plausibility models.

Definition 1.17 (Unified (Local) plausibility models [40, 16])**.** *A* unified or local plausibility model $\mathcal{M} = \langle W, (\trianglelefteq_i)_{i\in N}, V\rangle$ *has* $W \neq \emptyset$*, while, for each* $i \in N$*,* $\trianglelefteq_i$ *is a pre-order on* W *such that for each state* w*,* $\trianglelefteq$ *is connected and well-founded on each of its comparability classes* $\{v \mid v \trianglelefteq w \text{ or } w \trianglelefteq v\}$*, and* $V : \text{PROP} \to \wp(W)$*.*

For details about languages interpreted on such structures see [17, 40].

Now that we have presented epistemic plausibility models and indicated how the notions of (conditional) belief and knowledge are interpreted on them, we would like to introduce frameworks that bring these notions into the context of interactive systems of agents and study their evolution over time.

1.4 Global vs local models of rational agency

Some framework are concerned with the phenomenon of intelligent and rational interaction as a whole, taking into account the different dimensions of rational

behavior and reasoning involved in it. Game theory has such a global perspective, studying classes of predictable outcomes for classes of interactive situations. Other frameworks isolate certain sub-phenomena, certain aspects of intelligent agency and analyze them separately. Modal logics have been developed with the aim of determining the valid inference patterns on given classes of models representing a particular phenomenon of interest. Among them dynamic and temporal logics seem currently the two major logical paradigms concerned with some aspects of intelligent interaction.

Since these logical models are strongly connected with the models used by game theory, in its analysis of strategic interaction between "intelligent rational decision-makers" (Myerson [124]), we think it is interesting to go immediately with some detail into what game theory is about before moving on to our primary interest: *modal logics for reasoning about belief change*. Introducing game theory will give an idea of the use of global mathematical models of strategic interaction, taking into account agents' capacities, preferences, information and belief to analyze and understand how intelligent agents reason and make decisions in interactive contexts (Section 1.5). We can then see how the two important logical paradigms, the dynamic and the temporal approaches, deal with specific phenomena (Section 1.6), with a special interest for those focusing on information and belief dynamics.

1.5 Game theory

"Game theory analyses situations in which two or more individuals make decisions that will influence one another's welfare" [124]. It develops mathematical models that aim at explaining strategic multi-agent decision-making. A game is an abstract model of an interactive decision process. Game theory is interested in classes of reasonable outcomes for classes of games, or *solution concepts*. There are however different interpretations of the basic concepts of game theory, in particular of that of a solution concept. We think it is worth distinguishing between these different interpretations before proceeding further.

1.5.1 On the interpretation of games

There are at least three ways to interpret a game. They essentially differ on their interpretation of the concepts of *strategies*, *solution concepts* and — when they appeal to them — of *beliefs*. These interpretations can be called *epistemic*, *evolutive* and *evolutionary*. Let us immediately indicate that we won't be concerned with the third interpretation. Indeed it does not think of games as played by intelligent decision-makers but by populations of automata, that may or may not be evolutionary stable. In such an analysis, belief and belief change have obviously no place.

The epistemic interpretation of games.

In the epistemic perspective, games are one-shot events. Players will play against each other for the first time and they have no reason to expect the interactive situation to occur again. They may still have beliefs about the other players' strategies (first-order beliefs) and about other players' beliefs (higher-order beliefs). But these beliefs are not frequentist and do not result from previous play. Therefore *equilibrium* concepts, in the sense of long-term stability of profiles of beliefs and strategies (see *evolutive interpretation*), are *not* of primary interest. More generally for a solution concept to be interesting under the epistemic interpretation, it should be possible to give epistemic (and doxastic) *satisfiable* conditions under which players will play according to the solution concept. Such epistemic (or doxastic) conditions are called *epistemic foundations*. We will give an example of such a foundation in Section 1.5.2 when we have introduced the mathematical models.

The evolutive interpretation of games.

In the evolutive interpretation, players will play against each other more than once, learn from their mistakes, and adjust their behavior, i.e. their strategies. They can have beliefs about the other players' strategies formed by observation of previous play. At some point players might reach a state in which their expectations match the actual behavior of the other players and in which given the behavior of the other players they don't have any incentive to change their own strategy. They reached an *equilibrium*. Under the evolutive interpretation *equilibrium* concepts are of primary interest, Nash equilibrium being the most famous. Moreover some equilibria might be less fragile than others, e.g. with respect to out-of-equilibrium experimentation. Therefore different refinements are considered in the (evolutive) literature.

1.5.2 Games as models.

A *game* specifies the different choices that the players have, and their preferences over the outcomes of those (collective) choices. We will distinguish between games along three lines: strategic and extensive-form (or simply extensive) games, games with perfect information and games with imperfect information, and finally games with complete and with incomplete information.

A *strategic* game is a game in which all players make a single choice and make it simultaneously, while an *extensive-form* game is a game in which players make one or more decisions and make them sequentially. For the later kind of games players may or may not be perfectly informed about the sequence of actions that has been taken so far by the other players. Games in which every player is always certain of the previous history of actions when she is to move are said to be

extensive games with *perfect* information (while the others are said to be with *imperfect* information).

Finally players may or may not be completely informed about all parameters of the game, they might e.g. be uncertain about another player's preferences. In games of incomplete information outcomes are determined by the profile of actions taken by the agents *and by the state of the world*.

Let us only consider extensive games with incomplete information (strategic games of incomplete information won't play a role in this dissertation). In the probabilistic context Harsanyi [98] has shown that — under some assumptions — such a game could be equivalently analyzed as an extensive game with *imperfect* but complete information in which an additional agent Chance or Nature (without strategic interest) will draw according to some probability distribution the state of the world (e.g. the preferences of a player). About the actual draw, agents might receive more or less accurate information. Therefore we can focus on extensive games of imperfect information.

We first introduce strategic games of complete information and then turn to extensive games of imperfect information.

Definition 1.18 (Strategic game, see e.g. [127])**.** *A* strategic game $\mathcal{G}$ *based on a set of agents* N *is of the form* $((A_i)_{i\in N}, (\succeq_i)_{i\in N})$*. For each* $i \in N$*,* $A_i \neq \emptyset$ *is the set of actions available to* i*,* $\preceq_i$ *is a total pre-order on* $\times_{j\in N} A_j$ *(*i*'s preference relation). Given a strategy profile* $s \in \times_{i\in N} A_i$*, let* $s_{-i} = \times_{j\in N\setminus\{i\}} A_j$*.*

Let us mention two well-known solution concepts: pure Nash equilibrium and iterated strict dominance. A Nash equilibrium in pure strategies is a strategy profile that finds itself in a stable state in the sense that given the strategy chosen by the other players in a pure Nash equilibrium no player would have been better off by choosing another pure strategy. Another way to give the intuition is to say that if a player expects all others to conform to a Nash equilibrium then playing herself according to this Nash equilibrium is (one of) her best choice(s) or response(s). Formally:

Definition 1.19 (Nash equilibrium in pure strategies, see e.g. [127])**.** *A* Nash equilibrium in pure strategies *of a strategic game* $((A_i)_{i\in N}, (\succeq_i)_{i\in N})$ *is a strategy profile* $s^* \in \times_{i\in N} A_i$ *such that for every* $i \in N$ *and strategy* $a_i \in A_i$ *we have:* $(a^*_{-i}, a^*_i) \succeq_i (a^*_{-i}, a_i)$*.*

On the other hand, iterated strict dominance (or iterated elimination of strictly dominated actions) is an iterative procedure by which actions that can never be best responses, i.e. cannot be optimal for any profile whatsoever of strategies for the other agents, are iteratively eliminated.

Definition 1.20 (Strictly dominated actions)**.** *An action* a_i *is strictly dominated by an action* b_i *in a game* $((A_i)_{i\in N}, (\succeq_i)_{i\in N})$ *if for all profile strategies* s_{-i} *available for the other agents,* $(s_{-i}, b_i) \succ_i (s_{-i}, a_i)$*, and* a_i *is strictly dominated if it is dominated by some action.*

We give an informal definition of (surviving) iterated strict dominance and refer to e.g. [127, ch.4] for a formal definition.

Definition 1.21 (Surviving iterated strict dominance)**.** *Given a game, iterated strict dominance eliminates all actions that are strictly dominated, generating a reduced game, and restarts from this reduced game, until it reaches a fixed point. A profile of strategy survives iterated strict dominance if it belongs to the strategy space in the fixed point.*

To illustrate the epistemic approach, we give an example of an epistemic foundation of a solution concept that we state informally:

Theorem 1.22 (Tan and Werlang [150])**.** *An action a_i can rationally be chosen under common belief of rationality iff a_i survives iterated strict dominance.*

The first direction of the theorem says that if in some state of a doxastic model of the game it is common belief between the players that they will take an action that is a best response to their beliefs, then at this state they are choosing an action that survives iterated strict dominance.

We now turn to extensive-form games with imperfect information. As for other relations, we will write $<[t]$ to be the image of t under the relation $<$, i.e. $<[t] = \{s \in T \mid t < s\}$.

Definition 1.23 (Extensive-form game with imperfect information, see e.g. [124])**.** *An* extensive-form game *is of the form*

$$(T, <, Z, N, \rho, (\equiv_i)_{i\in N}, Act, A, (u_i)_{i\in N}), \text{ where}$$

- $(T, <)$ *is a finite rooted tree,*
- $Z = \{t \in T \mid <[t] = \emptyset\}$ *is the set of terminal nodes,*
- $\rho : T \setminus Z \to N \cup \{c\}$ *indicates which player, possibly Chance, is to move.*
- $\equiv_i$ *is an equivalence relation on $\rho^{-1}(i)$ encoding i's information. We write $\mathcal{K}_i[t] := \{s \ : t \equiv_i s\}$, for i's information cell at t.*
- *Act is a set of actions and for each (t, s) such that $t < s$, $A(t, s)$ is the particular action that would lead from t to s. (Formally $A : \{(t, s) \mid t < s\} \to Act$ is such that for every $t \in T$ the restriction of A to $\{t\} \times <[t]$ is an injection).*
- *Given some non-terminal node t, let $A[t] := \{A(t, s) \mid t < s\}$. A should be such that $A[t] = A[s]$ whenever $t \equiv_i s$, for some i.*
- *Finally $u_i : Z \to \mathbb{R}$ is player i's utility function.*

We will focus on extensive games with imperfect information "in which at every point every player remembers what he knew in the past" [127]. We call this notion game-theoretical perfect recall to differentiate it from a related and less demanding notion of perfect recall that we will meet later in the context of epistemic temporal logics. Let us first define the notion of *record*.

Definition 1.24 (Record of player i's experience; [127]). *The record of player i's experience at t, $X_i(t)$ for $t \in T$, is the sequence of information cells that the player i has encountered until t and the action she has taken at them.*

We now define the game-theoretical notion of perfect recall.

Definition 1.25 (Game-theoretical perfect recall; [127]). *And extensive-form game with imperfect information has the game-theoretical perfect recall property if for each agent i, $\mathcal{K}_i[t] = \mathcal{K}_i[t']$ implies $X_i(t) = X_i(t')$.*

As for strategic games different equilibrium notions and solution algorithms have been considered to take into account the sequential structure of a game. We will meet some of them in our discussion of strategic reasoning in Chapter 6.

This was a very compact and partial introduction to game theory. For a detailed presentation the reader is referred to one of the following textbooks: [127, 124, 126, 78].

Let us conclude this introductory section on the basics of game theory by mentioning that logical approaches to solution concepts and their epistemic foundations have been developed in recent years, notably by [7, 31, 40, 43, 21]. The logical systems considered in these papers combined ideas of the more local approaches offered by different temporal and dynamic logics — local in the sense that each of these logics isolates a particular phenomenon of interest. Such temporal and dynamic logics are the subject of our next section.

1.6 Dynamic and temporal logics for belief change

Dynamic and temporal logic offer two important and complementary perspectives on particular phenomena of interest in the context of an analysis of intelligent interaction:

- Temporal logics give a more global view on particular scenarios describing all possible evolution of some multi-agent system. Some of these temporal logics deal specifically with the evolution of agents' knowledge and beliefs.
- Dynamic logics consider epistemic events as generic entities and study the logics of such events. As an example the concept of a public announcement corresponds to a particular class of event models that can be applied to and transform any given multi-agent situation and the beliefs that agents entertain; and it is possible to give its logic.

There are naturally relations between the two approaches and making them explicit is something we will pay special attention to.

1.6.1 Dynamic logics

Dynamic logics of model change study different types of informational events and how they transform the informational and doxastic dimensions of multi-agent (social) situations. Two of such logical systems have played a foundational role: public announcement logic (PAL) and dynamic epistemic logic (DEL) (the latter being a generalization of the former). Both of these systems take epistemic models as their basis and investigate logically how such models evolve under new information taking the form of epistemic events.

PAL

The logic of public announcements or PAL [134, 83, 15] is concerned with how agents' information (what they know) and high-order information (what they know about each other's information) evolves as (epistemic) facts are publicly announced. The language of PAL — $\mathcal{L}_{PAL}$ — extends that of EL and includes modalities of the form $\langle !\varphi \rangle$, meaning '*after φ is (publically and truthfully) announced, ...*'. Given an epistemic model $\mathcal{M}$, let $\mathcal{M}|\varphi$ be its relativization to $||\varphi||^{\mathcal{M}} = \{w \in |\mathcal{M}| \mid \mathcal{M}, w \Vdash \varphi\}$. The truth condition for public announcements is then:

Definition 1.26 (Truth condition for public announcements)**.**

$$\mathcal{M}, w \Vdash \langle !\varphi \rangle \psi \quad \text{iff} \quad \mathcal{M}, w \Vdash \varphi \text{ and } \mathcal{M}|\varphi, w \Vdash \psi$$

Axiomatization. The set of formulas of $\mathcal{L}_{PAL}$ valid over the class of all epistemic models can be axiomatized by extending **EL** with the following axioms:

$(!p)$	$\vdash \langle !\varphi \rangle p \leftrightarrow (\varphi \wedge p)$
$(!\neg)$	$\vdash \langle !\varphi \rangle \neg \psi \leftrightarrow (\varphi \wedge \neg \langle !\varphi \rangle \psi)$
$(!\wedge)$	$\vdash \langle !\varphi \rangle (\psi \wedge \chi) \leftrightarrow (\langle !\varphi \rangle \psi \wedge \langle !\varphi \rangle \chi)$
$(!\hat{K}_i)$	$\vdash \langle !\varphi \rangle \hat{K}_i \psi \leftrightarrow (\varphi \wedge \hat{K}_i \langle !\varphi \rangle \psi)$

Table 1.3: Axiom system **PAL**.

Theorem 1.27 (Soudness of **PAL**, [134, 83, 15])**. PAL** *is sound with respect to the class of epistemic models.*

The soundness of **PAL** guarantees that a complete compositional analysis can be carried out into the epistemic language. Every formula of the public announcement language is thus equivalent to a formula of the epistemic language, whose validities are already decided by the axiom system **EL**.

Corollary 1.28 (see e.g. Blackburn et al. [39]). **PAL + EL** *is strongly complete with respect to the class of epistemic models.*

Dynamic epistemic logics and product update

DEL is a generalisation of public announcement logic (PAL) [134]. The seminal paper is Baltag et al. [20]. Building on that work, 'dynamic epistemic logic' (DEL) includes operators $\langle\alpha\rangle$, meaning '*after the event* α *occurs* ...'. The crucial idea is that, in the same way epistemic models encode static multi-agent situations, epistemic *event models* can encode (epistemic) events which transform the current informational multi-agent situation, i.e. transform the epistemic models.

Definition 1.29 (Event Models). *An* event model *is a triple:* $\epsilon = \langle E, (\sim_i^\epsilon)_{i\in N}, \mathtt{pre}\rangle$, *where* $E \neq \emptyset$ *is a set of events, for each agent* $i \in N$, $\sim_i^\epsilon$ *is a binary relation on* E, $\mathtt{pre} : E \to \mathcal{L}_{EL}$, *is a precondition function and* $\mathcal{L}_{EL}$ *is an epistemic language. A* pointed event model *is an event model with one distinguished element from* $|\mathcal{E}|$.

While event models encode different epistemic events, the new epistemic model encoding the new situation is computed according to a simple general mechanism: product update.

Definition 1.30 (Product Update). *The* product update *of an epistemic model* $\mathcal{M} = \langle W, (\sim_i)_{i\in N}, V\rangle$ *with an event model* $\epsilon = \langle E, (\sim_i^\epsilon)_{i\in N}, \mathtt{pre}\rangle$ *is the model* $\mathcal{M} \otimes \epsilon$ *whose states are the pairs* (w, e) *such that* w *satisfies the precondition of the event* e *and whose epistemic relations are defined as:*

$$(w, e) \sim_i' (w', e') \text{ iff } e \sim_i^\epsilon e', w \sim_i w'$$

and whose valuation is defined by

$$(w, e) \in V'(p) \text{ iff } w \in V(p), \text{for all } p \in \textsc{prop}.$$

An epistemic model describes what agents currently know, while product update creates the new epistemic situation after some informational event has taken place. Illustrations of the strength of this simple mechanism can be found in [15].

In adding these dynamic operators to static epistemic logic, DEL merges ideas from philosophy and computer science. [20] and [83] were seminal in the development of DEL. Operators matching epistemic events are interpreted as follows.

Definition 1.31 (Truth condition for epistemic actions modalities).

$$\mathcal{M}, w \Vdash \langle\epsilon, e\rangle\psi \quad \text{iff} \quad \mathcal{M}, w \Vdash \mathtt{pre}_\epsilon(e) \text{ and } \mathcal{M} \otimes \epsilon, (w, e) \Vdash \psi$$

The axiomatization in Table 1.4 is really a scheme that should be instantiated in the relevant ways. To each event model correspond a modality and a specific Action-Knowledge ($\langle\epsilon, e\rangle K_i$) axiom.

PL	$\vdash \varphi$, for all classical propositional tautologies φ
Nec	If $\vdash \varphi$, then $\vdash K_i\varphi$
$(\langle\epsilon, e\rangle p)$	$\vdash \langle\epsilon, e\rangle p \leftrightarrow (\mathtt{pre}_\epsilon(e) \wedge p)$
$(\langle\epsilon, e\rangle \neg)$	$\vdash \langle\epsilon, e\rangle \neg\psi \leftrightarrow (\mathtt{pre}_\epsilon(e) \wedge \neg\langle\epsilon, e\rangle\psi)$
$(\langle\epsilon, e\rangle \wedge)$	$\vdash \langle\epsilon, e\rangle(\psi \wedge \chi) \leftrightarrow (\langle\epsilon, e\rangle\psi \wedge \langle\epsilon, e\rangle\chi)$
$(\langle\epsilon, e\rangle \hat{K}_i)$	$\vdash \langle\epsilon, e\rangle \hat{K}_i\psi \leftrightarrow (\mathtt{pre}_\epsilon(e) \wedge \hat{K}_i \bigvee_{f:e\sim_i^\epsilon f} \langle\epsilon, f\rangle\psi)$

Table 1.4: Axiom system **DEL**.

Theorem 1.32 (Soudness of **DEL**, [20])**.** **DEL** *is sound with respect to the class of epistemic models.*

For the same reason as for **PAL**, completeness follows.

Corollary 1.33 (Baltag et al. [20])**.** **DEL** + **EL** *is strongly complete with respect to the class of epistemic models.*

Dynamic epistemic logics deal with strong signals, changing our information, removing uncertainties in a radical way, but other types of signals might affect our beliefs in much weaker ways. I might believe the news on the radio, but I might not take it for certainty as I would with some observation I made directly. The DEL approach can be extended to cases of belief change building on epistemic plausibility models. Such *dynamic logics for belief change* have been developed recently (van Benthem [29], Baltag and Smets [16]) considering the type of events that trigger changes not only in the uncertainty relation but also in the plausibility ordering. We decided to leave the introduction of dynamic doxastic logic (DDL), and the important mechanism of Priority Update, to the next chapter in Section 2.3. Similarly we will leave the introduction of doxastic temporal logics to that chapter. Indeed Chapter 2 will carry out a systematic comparison of these two logic-based approaches to belief change and we thought it was natural to present these two approaches together in detail in that chapter.

1.6.2 Temporal logics

As for temporal logics of agency, they can be divided in two categories, depending on which aspect of multi-agent systems the logics focus on: what agents can achieve or what agents know.

Temporal logics of agency

On the one hand are temporal logics that study models which capture what agents or groups of agents can achieve. Such logics build on *temporal logics* developed in the computer science and philosophy community (Stirling [149] and Hodkinson

and Reynolds [103] give surveys) and/or neighborhood semantics [58]. Among the most prominent of these logics of individual and coalitional agency are Coalition Logic [131], Alternating-time temporal logic [4], STIT [24] and Game Logic [132]. Models are usually temporal trees or forests together with some functions or relations encoding the power of the agents, or some variant of labelled-transition systems based on neighborhood semantics that can be unfolded into temporal trees or forests. We refer to [105] for a detailed overview of these different systems.

We will be especially concerned with the extension of such logics with preferences (such as e.g. [112, 1]) in Chapter 7.

Epistemic temporal logics

On the other hand are *epistemic temporal logics* that focus on the evolution of agents' knowledge and the effect of different protocols on this evolution. These logics are most frequently interpreted either on Epistemic Temporal Forests [130] or on Interpreted Systems [72]. While the second approach considers agents' internal states as structural primitives, the first approach takes uncertainty relations to be the primitives. In this dissertation we work with Epistemic Temporal Forests, but the two approaches are deeply related. Pacuit [128] shows that the logics developed on both types of structures form a coherent family. To be more specific Pacuit [128] shows that for natural epistemic temporal languages a formula is satisfiable in a pointed Epistemic Temporal Model iff it is satisfiable in an Interpreted System. In this dissertation we prefer to work with Epistemic Temporal Forests.

We now turn to epistemic temporal models, introduced by Parikh and Ramanujam [130] as a Grand Stage of unfolding informational events. In what follows, Σ^* is the set of finite sequences on any set Σ, which naturally forms a branching 'tree'.

Definition 1.34 (Epistemic Temporal Models)**.** *An* epistemic temporal model *('ETL model')* $\mathcal{H}$ *is a tuple* $\langle \Sigma, H, (\sim_i)_{i\in N}, V\rangle$ *with* Σ *a finite set of events, and* $H \subseteq \Sigma^*$ *closed under non-empty prefixes. For each* $i \in N$*,* $\sim_i$ *is a binary relation on* H*, and there is a valuation* $V : \text{PROP} \to \wp(H)$*.*

Here the set of histories H functions as a *protocol* defining all admissible trajectories of an informational multi-agent process. We refer to the information of agent i at h by $\mathcal{K}_i[h] = \{h' \in H \mid h \sim_i h'\}$.

For some applications we will consider epistemic temporal models allowing ω-sequences. Formally ω-epistemic temporal models are a generalization of epistemic temporal models in which the set of histories H is a subset of $\Sigma^* \cup \Sigma^\omega$. We will also consider particular cases of epistemic temporal models equipped with a set of initial states W and with $H \subseteq W \circ (\Sigma^*)$ where $\circ$ stands for concatenation. We refer to such models as W-epistemic temporal models. ω-W-epistemic tem-

poral models have then: $H \subseteq W \circ (\Sigma^* \cup \Sigma^\omega)$. Finally in the case of W-epistemic temporal models we use the following notation:

Definition 1.35 (Bundle of sequences associated with a state w)**.** *Let* $\mathbb{P} : s \mapsto (\{s\} \circ (\Sigma^* \cup \Sigma^\omega)) \cap H$ *for* $s \in W$*. Intuitively,* $\mathbb{P}(s)$ *is the* protocol *or* bundle *of sequences of events associated with* s*. We refer to the* $\langle W, \Sigma, H\rangle$*-part of an ETL model as the protocol this model is based on.*

These approaches have been extended to the analysis of belief change over time. Such *doxastic temporal logics* as introduced by Friedman and Halpern [77] and Bonanno [47] represent time globally as a bundle of possible histories where the beliefs of agents evolve as informational processes unfold. As for the dynamic case, we leave the introduction of doxastic temporal logics to Chapter 2 in order to keep the presentation of the dynamic and the temporal approaches to belief change together before proceeding to their systematic comparison.

Finally we mentioned that our analyses will not only be carried out at the level of structural primitives but also at the syntactic level. The next section discusses modal languages that are richer than the basic modal language and systematic criteria both to check whether certain types of reasoning can be carried out in certain languages and to compare languages in terms of their expressive power.

1.7 Richer languages for agency: comparing and evaluating their expressive power

Throughout the dissertation we will often compare alternative languages to reason about particular structures, with a special focus on definability issues, i.e. on whether a language is strong enough to distinguish between certain pointed models or to characterize certain classes of frames. Answering definability questions will help us to determine which languages are suitable to capture certain types of reasoning or whether we can hope to give a syntactic counterpart to semantic results in a given language.

As it will turn out basic languages such as the basic epistemic languages $\mathcal{L}_{EL}$ or the basic doxastic language $\mathcal{L}_{DOX}$ will be generally too weak for the type of reasoning we will encounter. For many applications it will be useful to draw on the resources of hybrid languages and boolean modal logics which represent intermediate languages between basic modal languages and a full first-order language. We will briefly introduce these two types of extended modal languages in the next subsection.

As for deciding questions relating to the expressivity of modal logics, such as definability questions, we can base ourselves on known invariance results that characterize the expressive power of known classes of logics. The reader can find some background on these invariance results and important operations and relations on models that we refer to throughout the dissertation in Appendix B.

We present two natural ways of extending basic modal languages that have been considered in the literature. The first idea is to add syntactic counterparts to primitive semantic notions, notably using nominals to refer to states in the language. Hybrid languages follow this line. The second idea is to have modalities not only for the primitive relations, but for their union, their intersection and other constructs.

Hybrid languages

Any modal language can be extended by adding a new kind of propositional letter: *nominals* $(i, j, k \ldots)$ that will serve as *names* for states. They can be used in this way by having the valuation function map a nominal to a singleton of the domain ($V : \text{NOM} \to \wp(W)$ with for all $i \in \text{NOM}$, $|V(i)| = 1$). On top of that, we can add satisfaction operators $@_i, @_j \ldots$, with "φ holds at the state named by i" as the intended meaning of $@_i\varphi$. Given a language $\mathcal{L}_{XYZ}(\tau)$ we refer to $\mathcal{H}_{XYZ}(@, \tau)$ as its hybrid extension. Furthermore we can extend this language with state variables $x, y, z, \ldots$ and a binder $\downarrow x., \downarrow y., \ldots$ which binds state variables to the current state. They are used in this way by interpreting formulas at a pointed model together with an assignment function $g : \text{SVAR} \to W$, mapping state variables to states (elements of the domain). "Bind (interpret) x to (be) the current state, now φ holds in the current state" is the intended meaning of $\downarrow x.\varphi$ and finally we have satisfaction operators for state variables $@_x, @_y, @_z, \ldots$ Given a language $\mathcal{L}_{XYZ}(\tau)$ we refer to $\mathcal{H}_{XYZ}(@, \downarrow, \tau)$ (resp. $\mathcal{H}_{XYZ}(\downarrow, \tau)$) as its hybrid extension with binders, state variables and with (resp. without) satisfaction operators.

Let us give a concrete example.

Definition 1.36 (Hybrid Epistemic Language with binders and satisfaction operators)**.** *The epistemic language $\mathcal{H}_{EL}(@, \downarrow, \tau)$ is defined as follows:*

$$\varphi ::= p \mid i \mid x \mid @_i\varphi \mid @_x\varphi \mid \downarrow x.\varphi \mid \neg\varphi \mid \varphi \wedge \varphi \mid K_j\varphi,$$

where j ranges over N, p over a countable set of proposition letters PROP*, i over a countable set of nominals* NOM*, and x over a countable set of state variables* SVAR*.*

$\mathcal{H}_{EL}(@, \downarrow, \tau)$ is interpreted on epistemic models together with an assignment $g : \text{SVAR} \to W$ as follows.

Definition 1.37 (Truth definition)**.**

$$\begin{array}{lll}
\mathcal{M}, w, g \Vdash p & \text{iff} & w \in V(p) \\
\mathcal{M}, w, g \Vdash i & \text{iff} & w \in V(i) \\
\mathcal{M}, w, g \Vdash x & \text{iff} & g(x) = w \\
\mathcal{M}, w, g \Vdash \downarrow x.\varphi & \text{iff} & \mathcal{M}, w, g[g(x) := w] \Vdash \varphi \\
\mathcal{M}, w, g \Vdash @_i\varphi & \text{iff} & \mathcal{M}, v, g \Vdash \varphi \text{ where } v \in V(i) \\
\mathcal{M}, w, g \Vdash @_x\varphi & \text{iff} & \mathcal{M}, g(x), g \Vdash \varphi \\
\mathcal{M}, w, g \Vdash \neg\varphi & \text{iff} & \mathcal{M}, w \nVdash \varphi \\
\mathcal{M}, w, g \Vdash \varphi \wedge \psi & \text{iff} & \mathcal{M}, w \Vdash \varphi \text{ and } \mathcal{M}, w \Vdash \psi \\
\mathcal{M}, w, g \Vdash K_i\varphi & \text{iff} & \text{for all } v \text{ such that } w \sim_i v \text{ we have } \mathcal{M}, v, g \Vdash \varphi
\end{array}$$

Boolean modal languages and PDL

A propositional letter is interpreted as a subset of the domain. And boolean connectives allow us to construct new formulas and to refer to a richer collection of subsets of the domain. A program is the syntactic counterpart to a relation in the model. For example we could think as K_i as scanning the program i which is interpreted as $\sim_i$. In the same way that new formulas can be constructed from propositional letters, boolean modal languages allow us to construct new programs from the basic programs according to some operations such as $\cup$, $\cap$ and complement. For some applications, it is enough to consider a few additional programs instead of taking the collection of allowed programs to be a complete algebra.

Let us take an example.

We can build a program which is interpreted as the intersection $\geq_i \cap \sim_i$. The necessity operator scanning this program is a new modality with its own properties. In fact the corresponding necessity operator is a weakly defeasible (S4)-knowledge operator of 'safe belief' [16] whose semantics is as follows:

$$\mathcal{M}, w \Vdash \Box_i\varphi \quad \text{iff} \quad \forall v \text{ with } v \leq_i w \text{ and } w \sim_i v \text{ we have } \mathcal{M}, v \Vdash \varphi$$

Given an arbitrary model we can go further and allow for the set of programs to be an algebra containing a set of atomic programs; a boolean algebra in the case of the boolean modal logic considered in Gargov and Passy [80], a Kleene algebra for PDL (see e.g. [69, 39]). But other combinations are naturally possible. By extension one often refers to logics defined in these ways as boolean modal logics or as propositional dynamic logics (often when containing iteration). As an example if we take the set of programs to be closed under intersection, union and iteration, our language will be defined as follows.

Syntax. Our language has a recursively defined set of programs:

$$\alpha ::= a \mid \alpha \cup \alpha \mid \alpha \cap \alpha \mid \alpha^*$$

where a ranges over a set of atomic programs A.

To each program correspond a modality $\langle\alpha\rangle$ in the language:

$$\varphi ::= p \mid \neg\varphi \mid \varphi \vee \varphi \mid \langle\alpha\rangle\varphi$$

where p ranges over proposition letters PROP.

Semantics. This language will be interpreted on models of the form $\langle W, (R_a)_{a\in A}, V\rangle$ as follows, starting with the interpretation of programs:

$$\begin{array}{lcl} R_a & = & R_a \\ R_{\alpha\cup\beta} & = & R_\alpha \cup R_\beta \\ R_{\alpha\cap\beta} & = & R_\alpha \cap R_\beta \\ R_{\alpha^*} & = & (R_\alpha)^* \end{array}$$

Now for the interpretation of (the new clause of) our language:

$$\mathcal{M}, w \Vdash \langle\alpha\rangle\varphi \quad \text{iff} \quad \text{for some } v \text{ with } wR_\alpha v \text{ we have } \mathcal{M}, v \Vdash \varphi$$

As an example, if we take our basic set of programs to correspond to the uncertainty relations for the agents, then common knowledge and distributed knowledge become definable as follows: $C_G\varphi \leftrightarrow [(\bigcup_{i\in G} \sim_i)^*]\varphi$ and $D_G\varphi \leftrightarrow [\bigcap_{i\in G} \sim_i]\varphi$.

As mentioned, extended modal languages will be worth considering to adjust the expressive power to our needs of definability, as driven by the applications we have in mind, i.e. the type of reasoning we would like to capture within the object-language. We will be interested in languages that are able to distinguish between certain situations as represented by relational models with a distinguished state (pointed models), but also in languages that are able to define relevant classes of frames. In both cases an important tool-box can help us to prove whether the definability task is feasible, before trying and defining crucial concepts explicitly. These tools are introduced in Appendix B.

1.8 Recapitulation and coda

We have now stated the general aims of this dissertation, put it in perspective among other fields, and provided some crucial existing background.

To summarize, this dissertation is concerned with intelligent agency, and in this context, it focuses on the information flow and belief dynamics that underlie it. Our aim is to give a unified logical perspective on belief change and the reasoning processes at work in intelligent interaction. Moreover, we want to use the resulting framework to link up with interactive reasoning (interactive epistemology), inductive reasoning (formal learning theory) and strategic reasoning (game theory). Having a logical approach allows us to extract patterns of reasoning common to all these different approaches. To do so we identify the right languages to define crucial notions and carry out the relevant reasoning.

Having a common perspective also allows us to compare assumptions across fields, and import ideas from one to another. To illustrate this, we will start by building an interface between dynamic logics of belief change and doxastic temporal logics at both the structural and the syntactic level, and from there reach out towards non-logical, sometimes even quantitative, approaches.

Summary and sources of the chapters

Chapter 2 is based on van Benthem and Dégremont [32].

This chapter builds bridges between the two prominent families of modal logics of belief change: dynamic doxastic logics computing stepwise updates, and temporal doxastic logics describing global system evolutions, both based on plausibility pre-orders. Following earlier results linking dynamic-epistemic and epistemic-temporal logics, we prove representation theorems showing under which conditions a doxastic temporal model can be represented as the stepwise evolution of a doxastic model under successive 'priority updates'. This then allows for merging, where, in particular, the notion of a 'temporal protocol' defining an informational process can be introduced into the more local dynamic perspective.

Chapter 3 is new, and it forms a natural syntactic counterpart to our first chapter. It studies formal languages for reasoning about multi-agent belief change, starting with static languages and their relative expressive power. It then moves to dynamic doxastic languages and their compositional analysis. Next we consider doxastic temporal languages, and definability issues of the important notions in the structural characterization of 'priority updaters' in the previous chapter. Finally, it gives a complete logic of protocol-based belief revision that exemplifies a merge of dynamic and temporal logics.

Chapter 4 is based on Dégremont and Roy [64].

In this chapter we bring Aumann's Agreement Theorem to dynamic-epistemic logic. We show that common *belief* of posteriors is sufficient for agreements in 'epistemic-plausibility models', under common and well-founded priors, from which the usual form of agreement results, using common knowledge, follows. We do not restrict to the finite case, and show that in countable structures such results hold if *and only if* the underlying 'plausibility ordering' is well-founded. We look at these results from a syntactic point of view, showing that neither well-foundedness nor common priors are expressible in a commonly used language, but that the static agreement result is finitely derivable in an extended modal logic. We finally consider 'dynamic' agreement results, show they have a counterpart in epistemic-plausibility models, and provide a new form of agreements via 'public announcements.' A comparison of the two types of dynamic agreements reveals that they can indeed be different.

Chapter 5 is based on Dégremont and Gierasimczuk [61].

This chapter studies the phenomenon of inductive reasoning from the interface between temporal and dynamic doxastic logics. It builds connections with formal learning theory, which formalizes the phenomenon of language acquisition — and can also be interpreted as a theory of empirical inquiry. The theory focuses on various properties of the process of *conjecture-change over time*. Treating 'conjectures' as beliefs, we link the process of conjecture-change to doxastic update. Using this approach, we reconstruct and analyze the temporal aspect of learning

in the context of temporal and dynamic logics of belief change. In particular, we translate learning scenarios into dynamic doxastic epistemic logic, and express finite identifiability as a problem of epistemic temporal logic model checking. Furthermore, we prove representation results of learnability conditions in terms of classes of doxastic epistemic temporal frames.

Chapter 6 builds on ideas from Dégremont and Zvesper [65].

This chapter focuses on strategic reasoning in extensive games of imperfect information. We distinguish and consider two families of logics that can be put to work to model strategic reasoning. These include temporal logics that are interpreted on structures that are almost the game itself, or a natural extension with additional relations. We also study dynamic logics that work with epistemic and doxastic models of games, and show how to model strategic reasoning as a model-changing operation. In both cases, we indicate with examples how the logics can be put to use to reason concretely, about key notions concerning players in strategic scenarios.

Chapter 7 is based on Dégremont and Kurzen [62].

This chapter is concerned with two other crucial dimensions of intelligent interaction: preferences, and coalitional group power. It studies expressivity and complexity of normal modal logics for reasoning about cooperation and preferences. We identify local and global notions for reasoning about cooperation of agents that have preferences. Many of these correspond to game- and social choice-theoretical concepts. We specify the expressive power required to express these by determining whether they are invariant under relevant operations on different classes of models and frames. We consider a large class of known extended modal languages, and show how the chosen notions can be expressed in well-chosen fragments. To determine how demanding reasoning about cooperation is in terms of computational complexity, we use known complexity results for extended modal logics, and obtain for each local notion an upper bound on the complexity of modal logics expressing it.

Chapter 8 summarizes the results of the dissertation, and states some larger and smaller open problems that arise when logic is used as a unifying medium in the way we have advocated here.

Chapter 2

Multi-agent belief change: Bridges between dynamic doxastic and doxastic temporal logics[1]

Our first task is to describe how agents revise their beliefs over time. As mentioned the phenomenon has been studied from many perspectives already, and our aim is not to increase the number of existing approches, but rather to unify them. We try to generate coherence by working with the methodology of dynamic epistemic (and dynamic doxastic) logic, but relating it systematically to other approaches, so that the apparent diversity in the field gets reduced, and connecting with yet further fields becomes easier (in later chapters). Since this is the broader framework for the whole dissertation, we devote quite some time to this theme, in fact, two chapters, one more semantic and one more syntactic.

2.1 Introduction

In this chapter we carry out a systematic comparison of two logic-based approaches at the structural level. One is *dynamic logics for belief change* that have been developed recently (van Benthem [29], Baltag and Smets [16]) using plausibility relations between worlds to represent agents' beliefs and conditional beliefs. An act of revision is then a single step of change in such a relation, triggered by some new incoming, hard or soft, information. Of course, such single steps can be iterated, leading to longer sequences. The other approach that we consider are *doxastic temporal logics* (cf. Halpern and Friedman [77], Bonanno [47]), representing time as a Grand Stage of possible histories where informational processes unfold.

In the process of comparing these frameworks, we do not operate in a void. Similar questions have been solved for knowledge in van Benthem and Pacuit [34],

[1] This chapter is based on van Benthem and Dégremont [32].

and van Benthem, Gerbrandy, Hoshi and Pacuit [36], in the form of representation theorems showing how sequences of models produced by 'product update' in dynamic-epistemic logic form a special subclass of epistemic temporal models in the sense of Fagin, Halpern, Moses and Vardi [72] and Parikh and Ramanujam [130]. In particular, these are the temporal models for agents endowed with Perfect Recall and 'No Miracles', learning by new observations only, possibly constrained by epistemic protocols. Our aim is to do the same for the dynamic doxastic logic of plausibility change by 'priority update', relating this to models of doxastic temporal logic. We will identify the crucial agent features behind dynamic doxastic belief revision, and position them inside the broader temporal setting. This is not just a simple generalization of the epistemic case, but the benefits are similar: comparability of frameworks, and interesting new research questions once they are merged. In this chapter, we concentrate on the representation aspect. Further development of the merged theory of dynamic agents in a doxastic temporal language and logic is found in the next chapter.

We start in the next section with basic terminology and background on earlier results for the epistemic setting. In Section 2.3.1 we motivate the choice of doxastic plausibility models as our representation of static multi-agent doxastic situations. We then present the *dynamic* step by step approach to belief change (Section 2.3), in particular, defining priority update. Next, the global *temporal* approach to beliefs over time is presented in Section 2.4. In Section 2.5 we show how step by step priority updates of a doxastic model, perhaps constrained by a protocol, generate a doxastic temporal model. The key temporal doxastic properties that characterize priority updaters are then identified and motivated in Section 2.6. In section 2.7 we prove our main result linking the temporal and dynamic frameworks, for the special case of *total* pre-orders, and then in general in Section 2.8. We discuss some variations and extensions in Section 2.9.

2.2 Background results

Epistemic temporal trees and dynamic logics with product update are complementary ways of looking at multi-agent information flow. Representation theorems linking both approaches were proposed for the first time in [28]. A nice presentation of these early results can be found in [119, ch5]. We briefly state a recent version from [36], referring the reader to that paper for a proof, as well as generalizations and variations.

We have defined epistemic models, event models and product update in Section 1.6.1 and epistemic temporal models, introduced by [130] as a Grand Stage of unfolding informational events, in Section 1.6.2.

While such *ETL* models are very general, many special constraints are possible. Some are the usual assumptions in epistemic logic, like having accessibility be an equivalence relation for $S5$-agents. But more important here are properties

connecting epistemic accessibility with flow of time, defining general properties of an informational process and the agents participating in it. Such agents can have more idealized or more bounded powers of observation, memory, and other cognitive features. In particular, the following epistemic temporal properties drive the main representation theorem in [36]:

Definition 2.1 (Basic Agent Properties)**.**

- **Perfect Recall** $\mathcal{H}$ *satisfies perfect recall iff* $\forall he, h'f \in H$ *if* $K_i[he] = K_i[h'f]$, *then* $K_i[h] = K_i[h']$. It states that agents do not forget past information as events take place.

- **Synchronicity** $\mathcal{H}$ *satisfies synchronicity iff* $\forall h, h' \in H$ *if* $K_i[h] = K_i[h']$, *then* $len[h] = len[h']$, *where* $len(x)$ *is the length of sequence* x. Synchronicity is satisfied if the agents have access to some external discrete clock and can thus keep track of the time.

- **Uniform No Miracles** $\mathcal{H}$ *satisfies uniform no miracles iff* $\forall h, h' \in H$ $\forall e_1, e_2 \in E$ *with* $he_1, h'e_2 \in H$, *if there are* $h'', h''' \in H$ *with* $h''e_1, h'''e_2 \in H$ *such that* $h''e_1 \sim_i h'''e_2$ *and* $h \sim_i h'$, *then* $he_1 \sim_i h'e_2$. Uniform no miracles characterizes agents that do not take into account the whole history but that proceed in a step by step way and only get new information by acts of observation.

- **Propositional stability** $\mathcal{H}$ *satisfies propositional stability iff for all* $h, he \in H$ *and* $p \in \text{PROP}$ *we have* $p \in V(he)$ *iff* $p \in V(h)$.

Dynamic-epistemic logic has borrowed one crucial idea from epistemic temporal logic. An *epistemic protocol* P maps states in an epistemic model to sets of finite sequences of pointed event models closed under taking prefixes. In general, this allows branching choices in a tree-like structure. This again defines the admissible runs of some informational process: not every observation may be available, or appropriate. More formally, let $\mathfrak{E}$ be the class of all *pointed event models*, having one 'actual event' marked. Then the set of protocols is $Prot(\mathfrak{E}) = \{\mathtt{P} \subseteq \mathfrak{E}^* \mid \mathtt{P} \text{ is closed under finite prefixes } \}$. Next comes the more general notion used in the recent literature:

Definition 2.2 (Local Protocols)**.** *Given an epistemic model* $\mathcal{M}$, *a local protocol for* $\mathcal{M}$ *is a function* $P : |\mathcal{M}| \to Prot(\mathfrak{E})$. *In the particular case where* P *is a constant function (mapping each world to the same set of sequences), we call the protocol* uniform. *Finally when the local protocol maps worlds to just a unique linear sequence of event models, we say that it is a* line protocol.

To avoid technicalities, in this chapter we state results with uniform line protocols. But our results generalize: see [36] for the epistemic case. Indeed, under

suitable renaming of events, making different event models disjoint, line protocols even have the same expressive power as general branching protocols.

Now, given an epistemic model $\mathcal{M}$ as our initial situation, plus a uniform protocol P, we can define the resulting temporal evolution as an epistemic-temporal model $Forest(M, P) = \bigcup_{\vec{\epsilon} \in P} M \otimes \vec{\epsilon}$, the 'epistemic forest generated by' $\mathcal{M}$ through sequential application of the pointed event models in P using product update $\otimes$.

Finally, we can state what iterated dynamic-epistemic update means in the broader setting of epistemic-temporal logic:

Theorem 2.3 (van Benthem et al. [36])**.** *Let $\mathcal{H}$ be an arbitrary epistemic-temporal ETL model. The following two assertions are equivalent:*

- *$\mathcal{H}$ is isomorphic to the temporal evolution $Forest(M, P)$ of some epistemic model $\mathcal{M}$ and uniform line protocol P,*
- *$\mathcal{H}$ satisfies Propositional Stability, Synchronicity, Bisimulation Invariance, Perfect Recall, and Uniform No Miracles.*

Thus, epistemic temporal conditions describing idealized epistemic agents characterize just those forests that arise from performing iterated product update governed by some protocol. [36] and [119, ch5] have details.

As stated in the introduction, our chapter extends this analysis to the richer setting of belief revision, where plausibility orders of agents evolve as they observe possibly surprising events. But to do so, we first need appropriate belief models, plus an appealing systematic revision mechanism.

Important remark about languages. Before moving on, it is important to stress one feature of the preceding representation theorem and results in its family. The precondition language for event models should exactly match the notion of *bisimulation*. This means that the language should be invariant under such bisimulations, and also, that it should be strong enough to characterize a pointed model up to such bisimulations. Two technical observations follow:

1. To get the right definability, we should either restrict attention to finitely branching *ETL* models (as in [36]), or alternatively, let the precondition function of product models take values in an infinitary epistemic logic.
2. These theorems can be parametrized, in the epistemic case, and even more so, the doxastic setting. We stay at a semantic level in this chapter, and state our results *up to language choice*. The next chapter discusses syntactic issues extensively, including other desiderata on the language, such as its expressive power for specifying the relevant properties of informational processes and the agents involved in them.

2.3 Dynamic logics of stepwise belief change (DDL)

Just like epistemic models, doxastic plausibility models change when appropriate triggering events are observed. It has become clear recently that a general mechanism for doing so works like the earlier product update ([16]).

2.3.1 Plausibility models: static doxastic situations

In this chapter we will focus on *doxastic plausibility models*, i.e. on pure (epistemic-free) plausibility models rather than *epistemic plausibility models*. The reason is that the general mechanism we are considering (Priority Update) to update these static situations takes care independently of the epistemic relations and of the plausibility ordering. Our analysis will also work for more complex structures. In fact it is very easy to extend our analysis to epistemic plausibility models by combining our results with the results for the epistemic case mentioned in the previous section. For the reasons mentioned in Section 1.3.3 we will state our results for both total and arbitrary pre-orders.

Remark: Alternatives. We have seen that some authors use models with just primitive plausibility relations. One can then define epistemic accessibility for a single agent as the union of that relation with its converse, accessing also less plausible worlds. We return to this perspective briefly in Section 2.9.3.

We must now consider how such models evolve as agents observe events.

2.3.2 Describing doxastic events

Let us now introduce the structures that describe complex doxastic events, crucially including the ways in which they appear to agents:

Definition 2.4 (Plausibility Event Model; [16])**.** *A* plausibility event model *('event model', for short)* ϵ *is a tuple* $\langle E, (\preceq_i)_{i\in N}, \mathtt{pre}\rangle$ *with* $E \neq \emptyset$*, each* $\preceq_i$ *is a pre-order on* E*, and* $\mathtt{pre} : E \to \mathcal{L}$*, where* $\mathcal{L}$ *is the basic doxastic language.*

As in the epistemic case, our analysis will work for various precondition languages for doxastic events. One specific choice is found at the end of Section 2.7. Combining perspectives, an 'epistemic plausibility event model' is a plausibility event model together with a collection of equivalence relations $(\sim_i)_{i\in N}$ on E.

In the following update rule, a new event itself comes with instructions as to how prior beliefs may be overridden. The principle is similar to that of 'Jeffrey [108] conditionalization' for probabilities: we follow the preferences of the plausibility event model, but if it leaves things open, we stick with prior preferences:

Definition 2.5 (Priority Update; [16])**.** Priority update *of a plausibility model* $\mathcal{M} = \langle W, (\preceq_i)_{i\in N}, V\rangle$ *and an event model* $\epsilon = \langle E, (\preceq_i)_{i\in N}, \mathtt{pre}\rangle$ *produces the plausibility model* $\mathcal{M} \otimes \epsilon = \langle W', (\preceq'_i)_{i\in N}, V'\rangle$ *defined as follows:*

- $W' = \{(w, e) \in W \times E \mid \mathcal{M}, w \Vdash \mathtt{pre}(e)\}$
- $(w, e) \preceq'_i (w', e')$ *iff either* $e \prec_i e'$, *or* $e \simeq_i e'$ *and* $w \preceq_i w'$
- $(w, e) \in V'(p)$ *iff* $w \in V(p)$, *for every* $p \in$ PROP

In the doxastic epistemic setting, priority update by an epistemic plausibility event model combines the preceding mechanism with product update, i.e. it has one more clause:

- $(w, e) \sim'_i (w', e')$ iff $w \sim_i w'$ and $e \sim_i e'$

More motivation for this rule can be found in [16], and at the end of this section. First here is a concrete example.

As mentioned, doxastic plausibility models are naturally combined with information partitions to describe scenarios involving both knowledge and beliefs. In this case priority update is applied to the plausibility ordering while product update is applied to the information partition. We will discuss this issue in connection with the temporal models in Section 2.9. Let us for now present a concrete scenario that involves both knowledge and beliefs.

Reading the figures. In the following figures, the actual state (respectively event taking place) is the shaded one. Epistemic equivalence classes are represented by rectangles or ellipses. We use $<$ to display the strict plausibility ordering within such classes. Our example assumes that all agents have the same plausibility ordering. An agent i believes φ at w is interpreted as φ holds in the i-most plausible states within i-information partition $K_i[w]$. An agent's beliefs at the actual state are thus displayed by an arrow from the actual state to the ones she considers most plausible, often just one. Thus, an arrow from x to y labelled by the agent *Enzo* means that y is the $\leq_e$-minimal state within $K_e[x]$. A similar convention applies to the event-model. Finally, we omit reflexive arrows throughout.

Example 2.6. ***Failed invitation.*** *Céline and Enzo would like to invite Denis to their Wii party. The party has been decided but none of them has informed Denis yet. Denis considers it a priori more plausible that no Wii party is taking place unless informed otherwise. This initial situation is common knowledge between Céline and Enzo. In the following figures, plain rectangles (or ellipses) will represent Denis' epistemic partition, dashed ones Enzo's and dotted ones Céline's;* w *and* $\overline{w}$ *are state names.*

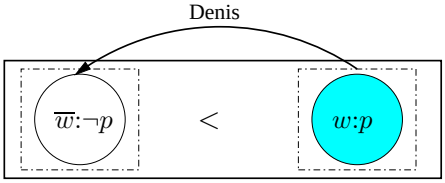

Figure 2.1: No Wii Party unless stated otherwise. Initial model.

The key event model. *The telephone rings and Céline picks up the phone. Enzo hears part of the conversation and concludes that Céline is inviting Denis. In fact Céline is not on the phone with Denis. Céline thinks it was clear from the conversation that she was not talking to Denis.* **e**, **f** *and* **g** *are event names.*

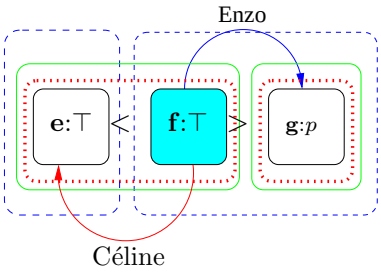

Figure 2.2: Event model of a misleading phone call.

We are now able to compute the new doxastic epistemic situation. The misunderstanding is now complete. In fact one can check that Enzo wrongly believes that it is now common knowledge between Céline and Denis that there is a Wii party while Céline wrongly believes that it is common belief between her and Enzo that Denis still does not know about the Wii party and even that Denis still believes that there is no Wii party.

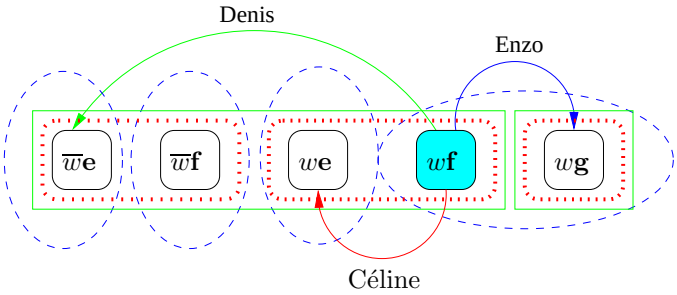

Figure 2.3: Product model of a misunderstanding.

Remark. Priority Update. In *AGM* style belief revision theory [3], new information is simply a new formula 'to be believed' by the agent. This allows for many different 'revision policies', from radical to conservative — a line also followed in a *DDL* setting by van Benthem [29]. It is important to appreciate that priority update is not just one such policy among many, but a *general mechanism* that can mimic many different policies depending on the richer structure of its triggers, viz. the plausibility event models [17]. If the event model has 'strong views', the update is radical, otherwise, it stays conservative. Interestingly, this mechanism shifts the variety in belief revision away from fixed agent types, to case-by-case decisions: I can be radical with one input, and conservative with another.

We feel that a logic should describe a 'universal' mechanism, instead of a jungle of styles. This is why we have chosen priority update, leading to one representation that covers all special cases.

2.4 Doxastic temporal models: the global view

We now turn to the temporal perspective on multi-agent belief revision, as an informational process over time with global long-term features. The following models are a natural doxastic enrichment of the temporal *ETL* models of [130]. They are also close to the temporal doxastic models of [47, 77]. First the doxastic temporal models:

Definition 2.7 (Doxastic Temporal Models)**.** *A* doxastic temporal model *('DoTL or DTL model' for short)* $\mathcal{H}$ *is of the form* $\langle \Sigma, H, (\leq_i)_{i\in N}, V\rangle$*, where* Σ *is a finite set of events,* $H \subseteq \Sigma^*$ *is closed under non-empty prefixes, for each* $i \in N$*,* $\leq_i$ *is a pre-order on* H*, and* $V : \text{PROP} \to \wp(H)$*.*

Doxastic Epistemic Temporal models (*DETL* models) are Doxastic Temporal models extended by a collection of epistemic equivalence relations $(\sim_i)_{i\in N}$ on H.

Given some history $h \in H$ and event $e \in \Sigma$, we let he stand for the concatenation of h with e. Given that plausibility links are not themselves events, the model H may again be viewed as a 'forest', a disjoint union of event trees. We sometimes refer to *DoTL* models as doxastic temporal forests. Figure 2.4 gives a concrete illustration of a practical setting with this abstract format. It display the evolution of a doctor's knowledge (dashed rectangles) and belief (diagnosis) — about what is wrong with her patient — as she performs medical tests and observes their positive or negative results (labelled edges). An arrow towards a state labelled `Environ` means that at this stage of the diagnostic process, the doctor think the patient's symptoms have an environmental cause. We omit reflexive and symmetric arrows.

Our models also gain concreteness by considering doxastic temporal languages interpreted on them. While these are the subject of the next chapter, we display a few truth conditions:

$$
\begin{array}{lll}
\mathcal{H}, h \Vdash \langle e\rangle\varphi & \text{iff} & \exists h' \in H \text{ with } h' = he \text{ and } \mathcal{H}, h' \Vdash \varphi \\
\mathcal{H}, h \Vdash \Box_i\varphi & \text{iff} & \forall h' \text{ with } h' \leq_i h \text{ and } h \sim_i h' \text{ we have } \mathcal{H}, h' \Vdash \varphi \\
\mathcal{H}, h \Vdash K_i\varphi & \text{iff} & \forall h' \text{ with } h \sim_i h' \text{ we have } \mathcal{H}, h' \Vdash \varphi \\
\mathcal{H}, h \Vdash B_i\varphi & \text{iff} & \forall h' \text{ with } h' \in \min_{\leq_i} K_i[h] \text{ we have } \mathcal{H}, h' \Vdash \varphi
\end{array}
$$

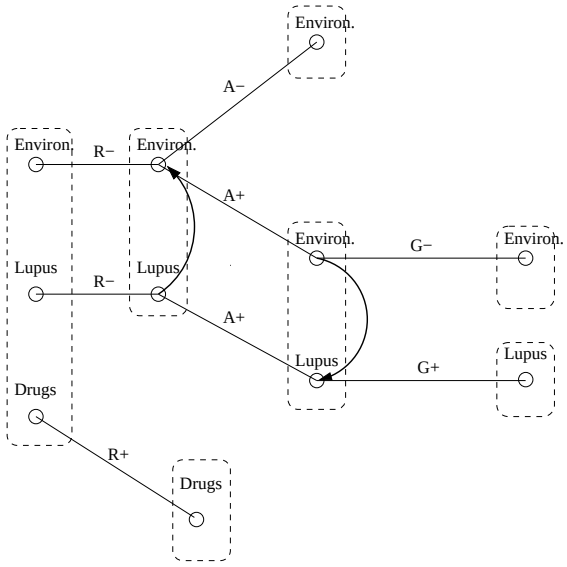

Figure 2.4: A medical investigation over time.

2.5 From DDL models to doxastic temporal models

Now we come to the main question of this chapter. Like *AGM*-style belief revision theory, Dynamic Doxastic Logic analyses one-step update scenarios. But, unlike *AGM* theory, it has no problem with iterating these updates to form longer sequences. Indeed let us put Example 2.6 together: Figure 2.5 looks like a doxastic epistemic forest model already. We will make this precise now, but as in the epistemic case, we need one more ingredient.

In many informational processes, such as learning, or belief revision in games, the information that agents receive may be highly constrained. Thus, there is crucial information in the set of admissible histories of the process, its 'protocol'. This notion can be defined formally just as before in Definition 2.2. Let $\mathfrak{E}$ be the class of all pointed plausibility event models. The set of *protocols* $Prot(\mathfrak{E}) = \{\mathtt{P} \subseteq \mathfrak{E}^* \mid \mathtt{P} \text{ is closed under finite prefixes}\}$. What we need is again a slightly more flexible version:

Definition 2.8 (Doxastic Protocols)**.** *Given a doxastic plausibility model $\mathcal{M}$, a local protocol for $\mathcal{M}$ is a function $P : |\mathcal{M}| \to Prot(\mathfrak{E})$. If P is a constant function, the protocol is called* uniform. *When P maps states to a linear nested sequence of event models, we call it a* line protocol.

We pointed out that the figure describing Example 2.6 really looks like a doxastic (epistemic) forest already. Actually we could continue the story, and the

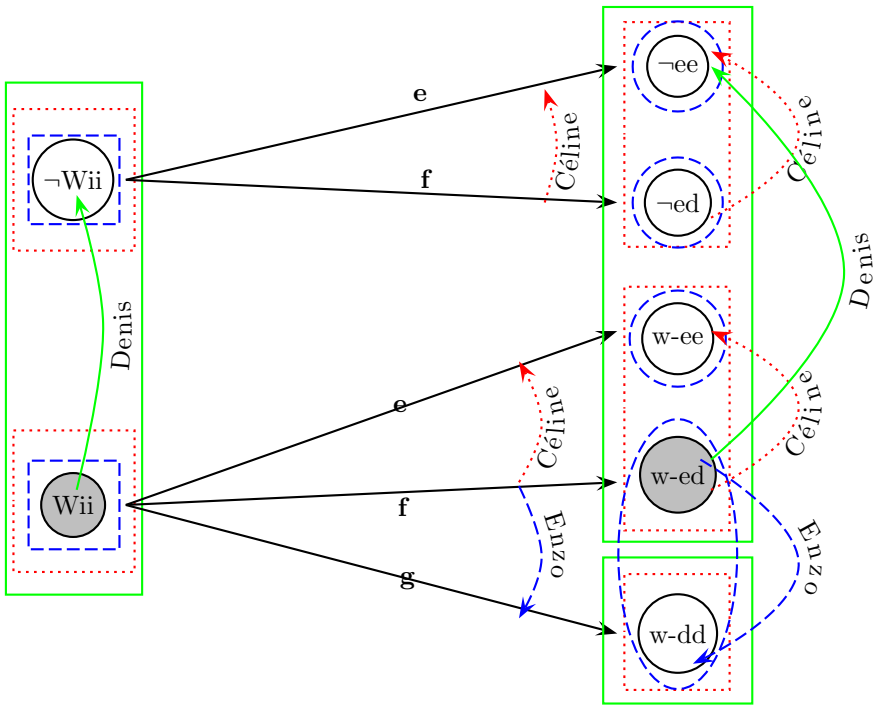

Figure 2.5: The Wii-party misunderstanding in temporal perspective.

further updates would generate a larger forest. More generally, priority update of a plausibility model according to a protocol generates a doxastic temporal forest.

In line with Section 2.2, we state our main theorems in terms of uniform line protocols. Iterated priority update of a doxastic plausibility model according to a uniform line protocol P generates a doxastic temporal forest model. We construct the forest by induction, starting with the doxastic plausibility model, and then checking which events can be executed according to the preconditions and to the protocol. Finally the new plausibility order is updated at each stage according to priority update. Since priority update describes purely doxastic, non-ontic change, the valuation stays the same as in the initial model. (For ways of adding real factual change, see [35].) For simplicity, we write $P(w) = \vec{\epsilon}$ where $\vec{\epsilon}$ is a finite sequence of event models.

Definition 2.9 ($DoTL$ model generated by a sequence of updates)**.** *Each initial plausibility model* $\mathcal{M} = \langle W, (\preceq_i)_{i\in N}, V\rangle$ *and each sequence of plausibility event models* $\vec{\epsilon} = (\epsilon_j)_{j\in\omega}$ *where* $\epsilon_j = \langle E_j, (\preceq_i^j)_{i\in N}, \mathtt{pre}_j\rangle$ *yields a* generated $DoTL$ plausibility model $\langle \Sigma, H, (\leq_i)_{i\in N}, \mathbf{V}\rangle$ *as follows:*

- *Let* $\Sigma := \bigcup_{i=1}^{m} E_i$*, with* $m = len(\vec{\epsilon})$*.*
- *Let* $H_1 := W$*, and for each* $1 < n \leq m$*, let* $H_{n+1} := \{(we_1 \dots e_n) \mid (we_1 \dots e_{n-1}) \in H_n$ and $\mathcal{M} \otimes \epsilon_1 \otimes \dots \otimes \epsilon_{n-1}, (we_1 \dots e_{n-1}) \Vdash \mathtt{pre}_n(e_n)\}$.

Finally let $H = \bigcup_{1 \leq k \leq m} H_k$.

- *If* $h, h' \in H_1$, *then* $h \leq_i h'$ *iff* $h \preceq_i^{\mathcal{M}} h'$.
- *For* $1 < k \leq m$, $he \leq_i h'e'$ *iff*
  1. $he, h'e' \in H_k$, *and*
  2. *either* $e \prec_i^k e'$, *or* $e \simeq_i^k e'$ *and* $h \leq_i h'$.
- *Finally, for all* $p \in$ PROP, *set* $wh \in \mathbf{V}(p)$ *iff* $w \in V(p)$.

Our task is to identify just when a doxastic temporal model is isomorphic to the 'forest' thus generated by a sequence of priority updates. In particular, this will uncover the key doxastic properties of agents assumed in this belief revision mechanism.

2.6 Crucial frame properties for priority update

We first get a few more general properties of our information process out of the way. The first of these merely says that in that process, the facts of the world do not change, only agents' beliefs about it:

Definition 2.10. *Let* $\mathcal{H} = \langle \Sigma, H, (\leq_i)_{i \in N}, V \rangle$ *be a DoTL model.* $\mathcal{H}$ *satisfies* propositional stability *if, whenever* h *is a finite prefix of* h', h *and* h' *satisfy the same proposition letters.*

Note that this can be generalized to include real world change. Next comes a basic property of the events that we allow as revision triggers:

2.6.1 Bisimulation invariance

The aim of this notion is to guarantee the existence of preconditions behind events in some modal language. Depending on the language parameter we choose, one has to choose the corresponding bisimulation notion. As mentioned in Section 2.2 we will state our results up to language choice, therefore we give an abstract definition of bisimulation below. We will however give a concrete example of language instantiation when stating a corollary of our result for doxastic epistemic models (Corollary 2.16 in Section 2.7). Let τ be a finite collection of binary relations $\langle R_1, \ldots, R_n \rangle$ on $H \times H$.

Definition 2.11 (τ-Bisimulation)**.** *Let* $\mathcal{H}$ *and* $\mathcal{H}'$ *be two DoTL-models based on the same alphabet* Σ. *A relation* $Z \subseteq H \times H'$ *is a* τ*-Bisimulation if, for all* $h \in H$, $h' \in H'$ *and all* $R_i \in \tau$

(prop) h *and* h' *satisfy the same proposition letters, whenever* hZh';

(back) If hZh' and hR_ij, then there is a $j' \in H'$ with jZj' and $h'R'_ij'$;

(forth) If hZh' and $h'R'_ij'$, then there is a $j \in H$ with jZj' and hR_ij.

If Z is a τ-bisimulation and hZh', we say h and h' are τ-bisimilar.

Definition 2.12 (τ-Bisimulation Invariance)**.** *A DoTL model $\mathcal{H}$ satisfies* τ-bisimulation invariance *if, for all τ-bisimilar histories $h, h' \in H$, and all events e, $h'e \in H$ iff $he \in H$.*

Note that these definitions apply also to *DETL* models. Here is an example. $(\sim_i \cap \leq_i)_{i \in N}$*-Bisimulation Invariance* will leave all formulas of the basic doxastic language with safe belief invariant, and hence our earlier preconditions for events. If we want these preconditions to be richer, then we need more clauses in the bisimulation — and the same is true if we want the bisimulation to preserve explicit temporal formulas involving events.

2.6.2 Agent-oriented properties

Now we come to the relevant agent properties. These depend on single agents i only, and hence we will drop agent labels and prefixes "for each $i \in N$" for the sake of clarity. Also, in what follows, when we write ha for events a, we assume that $ha \in H$.

Definition 2.13. *Let $\mathcal{H} = \langle \Sigma, H, (\leq_i)_{i \in N}, V \rangle$ be a DoTL model. $\mathcal{H}$ satisfies:*

- **Synchronicity** *Whenever $h \leq h'$, we have $len(h) = len(h')$.*

This says intuitively that agents have a correct belief about the exact stage the process is in. The following two properties trace the belief revising behavior of priority-updating agents more precisely:

- **Preference Propagation** *if $ja \leq j'b$, then $h \leq h'$ implies $ha \leq h'b$.*
- **Preference Revelation** *If $jb \leq j'a$, then $ha \leq h'b$ implies $h \leq h'$.*

What do the latter properties say? In the earlier epistemic representation theorems, the corresponding properties of Perfect Recall and No Miracles described observational agents with ideal memory, the two basic features behind the product update rule. Likewise, our new properties express the two basic features 'hard-wired into' the priority update rule, its 'radicalism' and its 'conservatism'. Preference Propagation says that, if the last-observed events ever allowed a plausibility preference, then they always do — or stated contrapositively, if they ever 'over-rule' an existing plausibility, then they always do. This reflects the first radical clause in the definition of priority update. Next, Preference Revelation says that when an agent has no strict plausibility preference induced by two observed events, then she will go with her prior plausibility. This reflects the second, conservative clause in priority update. As we have said before, this is a qualitative description of a 'Jeffrey-style' updating agent in a probabilistic setting.

2.7 The main representation theorem

Now we prove our main result relating DDL and DTL models, both with total pre-orders.

Theorem 2.14. *Let $\mathcal{H}$ be any doxastic-temporal model with a total plausibility pre-order. Then the following two assertions are equivalent:*

1. *There exists a total plausibility model $\mathcal{M}$ and a sequence of total plausibility event models $\vec{\epsilon}$ such that $\mathcal{H}$ is isomorphic to the forest generated by the priority update of $\mathcal{M}$ by the sequence $\vec{\epsilon}$.*
2. *$\mathcal{H}$ satisfies Propositional Stability, Synchronicity, Bisimulation Invariance, Preference Propagation, and Preference Revelation.*

Proof. **Necessity** $(1 \implies 2)$**.** We show that the given conditions are satisfied by any $DoTL$ model generated through successive priority updates along some given protocol sequence. Here, *Propositional Stability* and *Synchronicity* are straightforward from the definition of generated forests.

Preference Propagation. Assume that $ja \leq j'b$ (1). It follows from either clause in the definition of priority update that $a \leq b$ (2). Now assume that $h \leq h'$ (3). It follows from (2), (3) and again by priority update that $ha \leq h'b$.

Preference Revelation. Assume that $jb \leq j'a$ (1). It follows from the definition of priority update that $b \leq a$ (2). Now assume $ha \leq h'b$ (3). By the definition of priority update, (3) can happen in two ways. Case 1: $a < b$ (4). It follows from (4) by the definition of $<$ that $b \not\leq a$ (5). But (5) contradicts (2). We are therefore in Case 2: $a \simeq b$ (6), and so $h \leq h'$ (7).

Note that we did not make use of totality in this direction of the proof.

Sufficiency $(2 \implies 1)$**.** Given a $DoTL$ model $\mathcal{H}$ satisfying the stated conditions, we show how to construct a matching doxastic plausibility model and a sequence of event models.

Construction. Here is the initial plausibility model $\mathcal{M}_0 = \langle W, (\preceq_i)_{i \in N}, \hat{V} \rangle$:

- $W := \{h \in H \mid len(h) = 1\}$.
- Set $h \preceq_i h'$ iff $h \leq_i h'$.
- For every $p \in \text{PROP}$, $\hat{V}(p) = V(p) \cap W$.

Now we construct the j-th event model $\epsilon_j = \langle E_j, (\preceq_i^j)_{i \in N}, \mathtt{pre}_j \rangle$:

- $E_j := \{e \in \Sigma \mid$ there is a history $he \in H$ with $len(h) = j\}$.

- Set $a \preceq_i^j b$ iff there are $ha, h'b \in H$ such that $len(h) = len(h) = j$ and $ha \leq_i h'b$.
- For each $e \in E_j$, let $\texttt{pre}_j(e)$ be the formula that characterizes the set $\{h \mid he \in H \text{ and } len(h) = j\}$. By general modal logic, our condition of Bisimulation Invariance guarantees that there is such a formula. Again as mentioned at the end of Section 2.2 this sentence may be an infinitary one in general (if we don't assume the doxastic temporal models to be finitely branching). We give a concrete instantiation when we discuss the epistemic doxastic corollary of our result.

Now we show that the construction is correct in the following sense:

Claim 2.15 (Correctness)**.** *Let $\leq$ be the plausibility relation in the given doxastic temporal model. Let $\preccurlyeq^F_{DDL}$ be the plausibility relation in the forest model induced by priority update over the just constructed plausibility model $\mathcal{F}$ and the constructed sequence of event models. We have:*

$$h \leq h' \text{ iff } h \preccurlyeq^F_{DDL} h'.$$

Proof of the claim. The proof is by induction on the length of histories. The base case is obvious from the construction of our initial model $\mathcal{M}_0$. Now comes the induction step:

From *DoTL* **to** *Forest(DDL)***.** Assume that $h_1a \leq h_2b$ (1). It follows that in the constructed event model $a \leq b$ (2).

Case 1: $a < b$. By priority update we have $h_1a \preccurlyeq^F_{DDL} h_2b$, whatever relationship held between h_1 and h_2 in $\mathcal{F}$.

Case 2: $b \leq a$ (3). This means that there are h_3b, h_4a such that $h_3b \leq h_4a$. But then by *Preference Revelation* and (1) we have $h_1 \leq h_2$ in the original doxastic temporal model $\mathcal{M}$. It follows by the inductive hypothesis that $h_1 \preccurlyeq^F_{DDL} h_2$. But then, since a and b are indifferent by (2) and (3), priority update gives us $h_1a \preccurlyeq^F_{DDL} h_2b$.

From *Forest(DDL)* **to** *DoTL***.** Now let $h_1a \preccurlyeq^F_{DDL} h_2b$. Again we follow the two clauses in the definition of priority update:

Case 1: $a < b$. By definition, this implies that $b \not\leq a$. But then by the above construction, for all histories $h_3, h_4 \in H$ we have $h_3b \not\leq h_4a$. In particular we have $h_2b \not\leq h_1a$. But then by *totality* (this is the only place where we use this property), $h_1a \leq h_2b$.

Case 2: $a \simeq b$ (4) and $h_1 \preccurlyeq^F_{DDL} h_2$. For a start, by the inductive hypothesis, $h_1 \leq h_2$ (5). By (4) and our construction, there are h_3a, h_4b with $h_3a \leq h_4b$ (6). But then by *Preference Propagation*, (5) and (6) imply that we have $h_1a \leq h_2b$. QED

Remark. Corollary for the Doxastic Epistemic case. We get a representation result for the doxastic epistemic case as an immediate corollary from Theorem 2.14 and Theorem 2.3. Moreover we give a concrete instantiation of this corollary by choosing the language of Safe Belief. In the result below we refer to priority update as the result of applying product update to the epistemic relations and priority update to the plausibility orderings.

Corollary 2.16. *Let $\mathcal{H}$ be any doxastic epistemic temporal model with a total plausibility pre-order. Then the following two assertions are equivalent:*

1. *There exists a total epistemic plausibility model $\mathcal{M}$ and a sequence of total epistemic plausibility event models $\vec{\epsilon}$ taking preconditions in the modal language of Safe Belief such that $\mathcal{H}$ is isomorphic to the forest generated by the Priority Update of $\mathcal{M}$ by the sequence $\vec{\epsilon}$.*

2. *$\mathcal{H}$ satisfies Propositional Stability, Synchronicity, Perfect Recall, Uniform No Miracles, $(\sim_i \cap \leq_i)_{i\in N}$-Bisimulation Invariance, Preference Propagation, and Preference Revelation.*

Remark. It is naturally possible to product update *plausibility event models* by *epistemic event models* according to the following definition.

Definition 2.17 (Conservative product update)**.** *The (conservative)* product update *of epistemic plausibility model $\mathcal{M} = \langle W, (\sim_i)_{i\in\mathcal{A}}, \leq_i, V\rangle$ with an event model $\epsilon = \langle E, (\sim_i^\epsilon)_{i\in\mathcal{A}}, \mathtt{pre}\rangle$ is the model $\mathcal{M}\otimes\epsilon$ whose domain is $\{(w,e) \mid w \in W,\ e \in E \ \&\ \mathcal{M}, w \Vdash \mathtt{pre}(e)\}$. The epistemic relation in the resulting model is $(w,e) \sim_i' (w',e')$ iff $w \sim_i w'$ and $e \sim_i^\epsilon e'$, the plausibility ordering is $(w,e) \leq_i' (w',e')$ iff $w \leq_i w'$, and the valuation is as follows: $(w,e) \in V(p)$ iff $w \in V(p)$.*

But this would boil down to considering only epistemic signals rather than also including softer, genuinely doxastic types of incoming information. Indeed it is easy to see that such an update can always be simulated by priority update by extending the epistemic event model to a plausibility event model with a universal plausibility ordering on events, i.e. by defining $\leq_i^\epsilon = |\epsilon| \times |\epsilon|$. On the other hand the definition itself is really nothing more than *product update*. We will still sometimes refer to it as conservative product update to distinguish it from priority update. But let us return to the main issue.

The representation theorem we proved in this section (Theorem 2.14) shows how to find, inside the much broader class of all doxastic temporal models, those whose plausibility pattern was produced by a systematic priority update process.

2.8 Extension to arbitrary pre-orders

The preceding result generalizes to the general case of pre-orders, allowing incomparability. Here we need a new notion that was hidden so far:

Definition 2.18 (Accommodating Events)**.** *Two events $a, b \in \Sigma$ are pairwise* accommodating *if, for all $ga, g'b$: $(g \leq g' \leftrightarrow ga \leq g'b)$, i.e. a, b preserve and anti-preserve plausibility.*

We can now define our new condition on doxastic-temporal models:

- **Accommodation** Events a and b are accommodating in the sense of Def. 2.18 if both $ja \leq j'b$ and $ha \not\leq h'b$ for some j, j', h, h'.

Accommodation is a uniformity property saying that, if two events allow both plausibility orders for histories, then they are always 'neutral' for determining plausibility order. This property only comes into its own with pre-orders allowing incomparable situations:

Fact 2.19. *If $\leq$ is a total pre-order and $\mathcal{H}$ satisfies Preference Propagation and Preference Revelation, then $\mathcal{H}$ satisfies Accommodation.*

Proof. Assume that $ja \leq j'b$ (i) and $ha \not\leq h'b$. By totality, the latter implies $hb \leq h'a$ (ii). Now let $g \leq g'$. By Preference Propagation and (i), $ga \leq g'b$. Conversely, assume that $ga \leq g'b$. By Preference Revelation, (i) and (ii), we have $g' \leq g$. QED

We can also prove a partial converse without assuming totality:

Fact 2.20. *If $\mathcal{H}$ satisfies Accommodation, it satisfies Preference Propagation.*

Proof. Let $ja \leq j'b$ (1) and $h \leq h'$ (2). Assume that $ha \not\leq h'b$. Then by Accommodation, for every $ga, g'b$, $g \leq g' \leftrightarrow ga \leq g'b$. So, in particular, $h \leq h' \leftrightarrow ha \leq h'b$. But since $h \leq h'$, we get $ha \leq h'b$: a contradiction. QED

Finally, an easy counter-example shows that, even with $\leq$ total:

Fact 2.21. *Accommodation does not imply Preference Revelation.*

Proof. Take the simplest model where the following holds: $h'b \simeq ha \simeq j'a \simeq jb$ and $h' < h \simeq j' \simeq j$. QED

With arbitrary pre-orders we need to impose Accommodation:

Theorem 2.22. *Let $\mathcal{H}$ be any doxastic-temporal model with a plausibility pre-order. Then the following two assertions are equivalent:*

1. *There exists a plausibility model $\mathcal{M}$, and a sequence of plausibility event models $\vec{\epsilon}$ such that $\mathcal{H}$ is isomorphic to the forest generated by the Priority Update of $\mathcal{M}$ by the sequence $\vec{\epsilon}$.*
2. *$\mathcal{H}$ satisfies Bisimulation Invariance, Propositional Stability, Synchronicity, Preference Revelation, and Accommodation.*

By Fact 2.20, Accommodation also gives us Preference Propagation.

Proof. **Necessity of the conditions.** $(1 \implies 2)$ Checking the conditions in Section 7 did not use totality. So we focus on the new condition:

Accommodation. Assume that $ja \leq j'b$ (1). It follows by the definition of priority update that $a \leq b$ (2). Now let $ha \not\leq h'b$ (3). This implies by priority update that $a \not\leq b$ (4). By definition, (2) with (4) imply that $a \simeq b$ (5). Now assume that $g \leq g'$ (6). It follows from (5), (6) and priority update that $ga \leq g'b$. The other direction is similar.

Sufficiency of the conditions. (2 $\Longrightarrow$ 1) Given a *DoTL* model, we again construct a *DDL* plausibility model plus a sequence of event models:

Construction. The plausibility model $\mathcal{M}_0 = \langle W, (\preceq_i)_{i\in N}, \hat{V}\rangle$ is as follows:

- $W := \{h \in H \mid len(h) = 1\}$,
- Set $h \preceq_i h'$ whenever $h \leq_i h'$,
- For every $p \in$ PROP, $\hat{V}(p) = V(p) \cap W$.

We construct the j-th event model $\epsilon_j = \langle E_j, (\preceq_i^j)_{i\in N}, \mathtt{pre}_j\rangle$ as follows:

- $E_j := \{e \in \Sigma \mid$ there is a history of the form $he \in H$ with $len(h) = j\}$.
- For each $i \in N$, define $a \preceq_i^j b$ iff either (a) there are $ha, h'b \in H$ such that $len(h) = len(h) = j$ and $ha \leq_i h'b$, or (b) [*a new case*] a and b are accommodating, and we put $a \simeq b$ (i.e., both $a \leq b$ and $b \leq a$).
- For each $e \in E_j$, let $\mathtt{pre}_j(e)$ be the basic doxastic formula characterizing the set $\{h \mid he \in H$ and $len(h) = j\}$. Bisimulation Invariance guarantees that there is such a formula (maybe infinitary).

Again we show that the construction is correct in the following sense:

Claim 2.23 (Correctness)**.** *Let $\leq$ be the plausibility relation in the doxastic temporal model $\mathcal{M}$. Let $\preccurlyeq^F_{DDL}$ be the plausibility relation in the forest $\mathcal{F}$ induced by successive priority updates of the plausibility model by the sequence of event models we just constructed. We have:*

$$h \leq h' \text{ iff } h \preccurlyeq^F_{DDL} h'.$$

Proof of the claim. We proceed by induction on the length of histories. The base case is clear from our construction of the initial model $\mathcal{M}_0$. Now for the induction step, with the same simplified notation as earlier:

From *DoTL* **to** *Forest*(*DDL*)**.** We distinguish two cases.

Case 1: $ha \leq h'b, h \leq h'$**.** By the inductive hypothesis, $h \leq h'$ implies $h \preccurlyeq^F_{DDL} h'$ (1). Since $ha \leq h'b$, it follows by the construction that $a \leq b$ (2). Then, by (1), (2) and priority update, we get $ha \preccurlyeq^F_{DDL} h'b$.

Case 2: $ha \leq h'b, h \not\leq h'$**.** Clearly, then, a and b are not *accommodating* and thus the special clause has not been used to build the event model, though we do have $a \leq b$ (1). By the contrapositive of Preference Revelation, we also conclude that for all $ja, j'b \in H$, we have $j'b \not\leq ja$ (2). Therefore, our construction gives $b \not\leq a$ (3), and we conclude that $a < b$ (4). But then by priority update, we get $ha \preccurlyeq^F_{DDL} h'b$.

From $Forest(DDL)$ **to** $DoTL$**.** We again distinguish two cases.

Case 1: $ha \preccurlyeq^F_{DDL} h'b, h \preccurlyeq^F_{DDL} h'$**.** By the definition of priority update, $ha \preccurlyeq^F_{DDL} h'b$ implies that $a \leq b$ (1). There are two possibilities.

Case 1.1: The special clause of the construction has been used, and a, b are accommodating (2). By the inductive hypothesis, $h \preccurlyeq^F_{DDL} h'$ implies $h \leq h'$ (3). But (2) and (3) imply that $ha \leq h'b$.

Case 1.2: Clause (1) holds because for some $ja, j'b \in H$ in the $DoTL$ model, $ja \leq j'b$ (4). By the inductive hypothesis, $h \preccurlyeq^F_{DDL} h'$ implies $h \leq h'$ (5). Now it follows from (4), (5) and Preference Propagation that $ha \leq h'b$.

Case 2: $ha \preccurlyeq^F_{DDL} h'b, h \not\preccurlyeq^F_{DDL} h'$**.** Here is where we put our new accommodation clause to work. Let us label our assertions: $h \not\preccurlyeq^F_{DDL} h'$ (1) and $ha \preccurlyeq^F_{DDL} h'b$ (2). It follows from (1) and (2) by the definition of priority update that $a < b$ (3), and hence by definition, $b \not\leq a$ (4). Clearly, a and b are not accommodating (5): for otherwise, we would have had $a \simeq b$, and hence $b \leq a$, contradicting (4).

Therefore, (3) implies that there are $ja, j'b \in H$ with $ja \leq j'b$ (6). Now assume for a contradiction that (in the $DoTL$ model) $ha \not\leq h'b$ (7). It follows from (6) and (7) by Accommodation that a and b *are* accommodating, contradicting (5). Thus we must have $ha \leq h'b$. QED

Given a doxastic temporal model describing the evolution of the beliefs of a group of agents, we have determined whether it could have been generated by successive 'local' priority updates of an initial plausibility model.

2.9 Additional extensions and variations of the theorem

Several further scenarios can be treated in the same manner. In particular, it is easy to combine the epistemic analysis in Section 2.2 with ours to include agents having both *knowledge and belief*. Here are three more directions:

2.9.1 From uniform to local protocols

So far we have considered uniform line protocols. We have already suggested that line protocols are powerful enough to mimic branching protocols through renaming of events, and then taking a disjoint union of all branching alternatives. But uniformity is a real restriction, and it can be lifted. *Local protocols* allow the set of executable sequences of pointed event models forming our current informational process to vary from state to state. Indeed, agents need not even know which protocol is running. As was done in [36] for the epistemic case, we can still get our representation theorems to cover this case, by merely dropping the condition of Bisimulation Invariance. While this seems a simple move, local protocols drastically change the complete dynamic-doxastic logic of the system.

2.9.2 Languages and bisimulations

As we have noted in Section 2.4, our doxastic-temporal models support various languages and logics. These will be pursued in the next chapter, but we do make a few points here. In our setting a doxastic-temporal language has two main purposes: (a) stating 'local' preconditions for events, (b) specifying 'global' properties of the temporal evolution of the current process. As is well-known [39] a choice of language here corresponds to a choice of a semantic invariance relation, usually some weaker or stronger variant of *bisimulation*. For instance, we have seen that if the precondition language contains a safe belief operator scanning the *intersection* of (the converse of) a plausibility $\leq_i$ relation and an epistemic indistinguishability relation $\sim$, then the *back* and *forth* clauses should not only apply to $\geq_i$ and $\sim_i$ separately, but also to $\geq_i \cap \sim_i$. (Indeed $\cap$ is not safe for bisimulation.) But this can be varied, and one can also have stronger notions of bisimulation, respecting more structure, that work for more expressive doxastic languages. And things get even more complicated if we allow temporal operators in our languages (cf. [36]). We do not want to commit to any specific choice here, since the choice of a language seems orthogonal to our main concerns in this chapter. We will discuss formal languages in the next chapter, taking definability of our major structural constraints as a guide.

2.9.3 Alternative model classes

We mentioned in Section 1.3.3 that one can also work with a primitive plausibility relation that merges epistemic indistinguishability and doxastic plausibility. A corresponding (priority) update rule is considered in [16], and we indicate briefly the notions involved in the corresponding representation result (Appendix C):

Definition 2.24 (Local Priority Update)**.** *The* Priority Update *of a* unified *plausibility model* $\mathcal{M} = \langle W, (\trianglelefteq_i)_{i\in N}, V\rangle$ *and a* $\trianglelefteq$*-event model* $\epsilon = \langle E, (\trianglelefteq_i)_{i\in N}, \mathtt{pre}\rangle$ *is the unified plausibility model* $\mathcal{M} \otimes \epsilon = \langle W', (\trianglelefteq'_i)_{i\in N}, V'\rangle$ *constructed as follows:*

- $W' = \{(w, e) \in W \times E \mid \mathcal{M}, w \Vdash \mathtt{pre}(e)\}$,
- $(w, a) \trianglelefteq'_i (w', b)$ *iff either 1.* $a \trianglelefteq_i b$, $b \not\trianglelefteq a$ *and* $w \trianglelefteq w' \vee w' \trianglelefteq w$ *or 2.* $a \trianglelefteq_i b$, $b \trianglelefteq a$ *and* $w \trianglelefteq w'$,
- $(w, e) \in V'(p)$ *iff* $w \in V(p)$ *for every* $p \in$ PROP.

We refer to this operation as *Local Priority Update*.
Here are our basic temporal doxastic agent properties in this setting:

- $\trianglelefteq$-**Perfect Recall** If $ha \trianglelefteq h'b$ we have $h \trianglelefteq h' \vee h' \trianglelefteq h$.
- $\trianglelefteq$-**Preference Propagation** If $h \trianglelefteq h'$ and $ja \trianglelefteq j'b$ then also $ha \trianglelefteq h'b$.
- $\trianglelefteq$-**Preference Revelation** If $ha \trianglelefteq h'b \wedge jb \trianglelefteq j'a$, also $h \trianglelefteq h'$.
- $\trianglelefteq$-**Accommodation** If ($ja \trianglelefteq j'b$, $h' \trianglelefteq h$ and $ha \not\trianglelefteq h'b$), for all $ga, g'b \in H$ ($g \trianglelefteq g' \leftrightarrow ga \trianglelefteq g'b$), and for all $g'a, gb \in H$ ($g \trianglelefteq g' \leftrightarrow gb \trianglelefteq g'a$).

In Appendix C we show how these conditions drive a general representation theorem similar to the one in Section 2.7 and 2.8.

2.10 Conclusion

Agents that update their knowledge and revise their beliefs leave an epistemic and doxastic 'trace' over time of epistemic and doxastic relations. We have determined the special constraints that capture agents operating with the 'local updates' of dynamic doxastic logic.

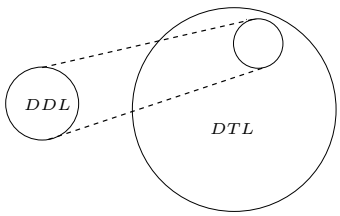

Figure 2.6: DDL inside DTL

Major sources. The first point of departure of this chapter is found in the recent work on dynamic logics of belief change by van Benthem [29], Board [40] and chiefly the sequence of papers by Baltag and Smets, in particular [16], who extend the dynamic epistemic methodology developed by Baltag et al. [20], Gerbrandy [83], Baltag and Moss [15] to belief change. The second source is constituted by the works on epistemic temporal logics by Parikh and Ramanujam [130], Fagin

et al. [72] and on doxastic temporal logics in a sequence of papers by Bonanno, in particular [47]. Finally the third source (but the most decisive) is van Benthem et al. [36], which carries out a systematic comparison between dynamic epistemic and epistemic temporal logic using the concept of protocol.

Our main results. Our contribution in this chapter was to extend the analysis of van Benthem et al. [36] to the case of belief change. This took the form of representation theorems that state just when a general doxastic temporal model is equivalent to the forest model generated by successive priority updates of an initial doxastic model by a protocol sequence of event models.

The next step. Thus we have determined the area where the idealized belief changers of dynamic doxastic logic live. Now that we have the contours of the semantics of belief revising agents over time, our next task is to bring out some of its essential features in a logical language, making them transparent to inspection, manipulation, and modification. That will be the task of our next chapter. We will discuss different dynamic doxastic and doxastic temporal languages, their expressive power, and the next chapter will make the preceding identification even stronger through an axiomatization of a temporal logic of belief revision.

Chapter 3

Merging modal logics of belief change: languages and logics

The previous chapter compared on the model-theoretic level the dynamic approach to belief change to the temporal one. It gave the structural foundations for reasoning about stepwise belief-revising agents over time. In this chapter we determine logical languages with the right expressive power to describe key features of belief revision agents, but also to enable reasoning about them. Our techniques are three major ones from modal logic: invariance, correspondence and completeness. We start by considering static doxastic languages and then move to the related dynamic doxastic and temporal doxastic languages. Finally we introduce, and prove completeness for, a temporal logic of belief revision.

3.1 Epistemic doxastic languages

We start with the epistemic doxastic languages which are 'static' languages. As such they are not saying anything about belief change but they give the foundations for both the temporal and the dynamic approach as they introduce the languages to reason about the doxastic and epistemic dimension of a given social situation. After recalling the clauses for knowledge and conditional belief, we consider other settings, including a language that matches closely the structural primitives of the now familiar *epistemic plausibility models*, on which all these languages will be interpreted.

We have introduced the basic epistemic doxastic language $\mathcal{L}_{DOX}$ in Section 1.3.3, that has modalities for both conditional beliefs and knowledge. Let us recall the important clauses:

$$\begin{array}{lll} \mathcal{M}, w \Vdash K_i\varphi & \text{iff} & \text{for all } v \text{ such that } w \sim_i v \text{ we have } \mathcal{M}, v \Vdash \varphi \\ \mathcal{M}, w \Vdash B_i^{\psi}\varphi & \text{iff} & \text{for all } v \text{ such that } w \rightarrow_i^{||\psi||^{\mathcal{M}}} v \text{ we have } \mathcal{M}, v \Vdash \varphi \end{array}$$

Other syntactic options are available, such as a language that matches closely the structural primitives of epistemic plausibility models, with the following syntax:

Definition 3.1. *The language $\mathcal{L}_{DOX}(\langle\geq_i\rangle, K_i)$ is defined as follows:*

$$\varphi ::= p \mid \neg\varphi \mid \varphi \wedge \varphi \mid K_i\varphi \mid \langle\geq_i\rangle\varphi,$$

where i ranges over N and p over a countable set of proposition letters PROP.

$\langle\geq_i\rangle\varphi$ reads 'there is a (a priori) more plausible state in which φ holds'. From the point of view of modal logic, $\langle\geq_i\rangle\varphi$ is a very natural operator, scanning the converse of the plausibility ordering. Interpreted on epistemic plausibility models, its truth condition is the obvious one:

$$\mathcal{M}, w \Vdash \langle\geq_i\rangle\varphi \quad \text{iff} \quad \text{for some } v \text{ with } v \leq_i w \text{ we have } \mathcal{M}, v \Vdash \varphi.$$

On the conceptual side, a sentence of the form 'agent i believes that φ' seems more intuitive than one of the form 'there is a state that i finds more plausible where φ holds'. But if the second language can simulate belief modalities, despite using these less intuitive modalities, it is still able to express the central notion of belief while remaining close to our structural primitives. So how do these languages relate in terms of expressive power? Maybe surprisingly, they are incomparable. Let us make the notion of comparability in expressive power precise.

Definition 3.2 (At least as expressive as)**.** *A language $\mathcal{L}_1$ is at least as expressive as a language $\mathcal{L}_2$, written $\mathcal{L}_2 \leq \mathcal{L}_1$ with respect to a class of pointed models $\mathbb{A}$ iff for every formula $\varphi \in \mathcal{L}_2$ there is a formula $\tau(\varphi) \in \mathcal{L}_1$ such that for every $\mathcal{M}, w \in \mathbb{A}$ we have $\mathcal{M}, w \Vdash \varphi$ iff $\mathcal{M}, w \Vdash \tau(\varphi)$.*

Let $\mathcal{L}_{DOX}(\langle\geq_i\rangle)$ be the K_i-free fragment of $\mathcal{L}_{DOX}(\langle\geq_i\rangle, K_i)$. We first show that $\langle\geq_i\rangle$ cannot be simulated in the basic doxastic language $\mathcal{L}_{DOX}$.

Fact 3.3. $\mathcal{L}_{DOX}(\langle\geq_i\rangle) \not\leq \mathcal{L}_{DOX}$.

Proof. Consider the upper and lower models in Figure 3.1. The basic doxastic language $\mathcal{L}_{DOX}$ cannot distinguish $\mathcal{M}, w$ from $\mathcal{M}', w'$ while $\mathcal{L}_{DOX}(\langle\geq_i\rangle)$ can.

That $\mathcal{L}_{DOX}(\langle\geq_i\rangle)$ can is obvious: $\mathcal{M}, w \Vdash \langle\geq_i\rangle\neg p$ but $\mathcal{M}', w' \nVdash \langle\geq_i\rangle\neg p$. That $\mathcal{L}_{DOX}$ cannot distinguish between $\mathcal{M}, w$ and $\mathcal{M}', w'$ can proved by induction. Propositional letters is by construction and the induction step for booleans is straightforward. For knowledge simply consider that the two states in the upper models are in two disjoint information cell. Now assume that we have completed the induction for formulas of degree n. Now consider the formula $B^\chi\psi$ with χ and ψ of degree n. Case 1: $w \in ||\chi||$ but then by IH so is w'. But then $\mathcal{M}, w \Vdash B^\chi\psi$ iff $w \in ||\psi||^{\mathcal{M}}$. Similarly $\mathcal{M}', w' \Vdash B^\chi\psi$ iff $w' \in ||\psi||^{\mathcal{M}'}$. But by IH we have $w \in ||\psi||^{\mathcal{M}}$ iff $w' \in ||\psi||^{\mathcal{M}'}$. Case 2: $w \notin ||\chi||$ but then $w' \notin ||\chi||$ and $B^\chi\psi$ is trivially true in both models. QED

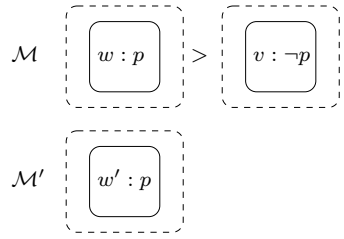

Figure 3.1: $\mathcal{L}_{DOX}$ is not as expressive as $\mathcal{L}_{DOX}(\langle \geq_i \rangle)$.

We now show that the basic epistemic doxastic language $\mathcal{L}_{DOX}$ cannot be simulated in $\mathcal{L}_{DOX}(\langle \geq_i \rangle, K_i)$.

Fact 3.4. $\mathcal{L}_{DOX} \not\leq \mathcal{L}_{DOX}(\langle \geq_i \rangle, K_i)$

Proof. Consider the upper and the lower model in Figure 3.2.

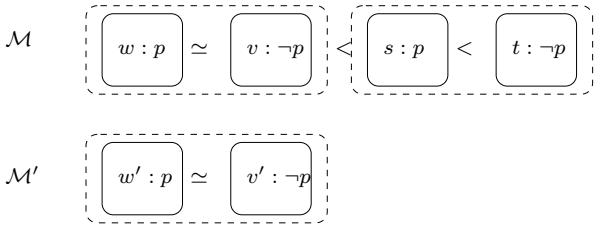

Figure 3.2: $\mathcal{L}_{DOX}(\langle \geq_i \rangle, K_i)$ is not as expressive as $\mathcal{L}_{DOX}$

We first show that $\mathcal{M}, t$ and $\mathcal{M}', v'$ are $\geq, \sim$-bisimilar. The witness bisimulation we are using is $Z = \{(w, w'), (v, v'), (s, w'), (t, v')\}$. Atomic harmony is immediate. That on the one hand $\mathcal{M}, w$ and $\mathcal{M}', w'$, and on the other hand $\mathcal{M}, v$ and $\mathcal{M}', v'$ are $\geq, \sim$-bisimilar is easy to check. It remains to prove the back and forth conditions (s, w') and (t, v').

By exploration of the first model we have $s \sim t$ but we also have $w' \sim v'$ and $(t, v') \in Z$. We also have $s \sim s$ but we have $w' \sim w'$ and $(s, w') \in Z$ so the forth condition for $\sim$ is satisfied. For the back condition we can proceed in a symmetric way. Now for $\geq$. We have $s \geq v$ but we also have $w' \geq v'$ and vZv'. We have $s \geq s$ but we also have $w' \geq w'$ and sZw'. For the other direction we have $w' \geq w'$ but then we have $s \geq s$ and sZw', and we have $w' \geq v'$ but then we have $s \geq v$ and $vZ'v'$. Finally we have $t \geq s$ but we also have $v' \geq w'$ and sZw'. We have $t \geq t$ but we also have $v' \geq v'$ and tZv'. For the other direction we have $v' \geq v'$ but then we have $t \geq t$ and tZv', and we have $v' \geq w'$ but then we have $t \geq s$ and $sZ'w'$.

Finally, that $\mathcal{L}_{DOX}$ can distinguish between $\mathcal{M}, t$ and $\mathcal{M}', v'$ is easy to see. Indeed $\mathcal{M}, t \Vdash Bp$, while $\mathcal{M}', v' \Vdash \neg Bp$ QED

What about further languages and further modalities? An intermediate notion of knowledge has been considered by Stalnaker [148] and by researchers in AI, and has been argued for doxastically as *safe belief* by [16] as describing those beliefs we do not give up under true new information, giving it a strong connection with ideas developed by Board [40] and, a few decades before, by formal epistemologists in the aftermath of the Gettier [84] problem (Baltag and Smets [17, 2.3] have pointers). Note that safe belief is only safe for *true* information. In contrast to knowledge, safe belief *can* be defeated by *false* information and in contrast to both knowledge and belief, safe belief is not negatively introspective (but it is still positively introspective). Formally, the safe belief modality $\Box_i$ is just the universal dual of the existential modality $\langle i \cap \geq_i \rangle$. Their truth conditions are:

$$\begin{array}{lll} \mathcal{M}, w \Vdash \langle i \cap \geq_i \rangle \varphi & \text{iff} & \text{for some } v \text{ with } v \leq_i w \ \& \ w \sim_i v \text{ we have } \mathcal{M}, v \Vdash \varphi \\ \mathcal{M}, w \Vdash \Box_i \varphi & \text{iff} & \text{for all } v \text{ such that } v \leq_i w \ \& \ w \sim_i v \text{ we have } \mathcal{M}, v \Vdash \varphi \end{array}$$

Interestingly together with knowledge, the safe belief modality supplies us with the right expressive power to simulate the basic doxastic language. To see that, let $\mathcal{L}_{DOX}(K_i, \langle i \cap \geq_i \rangle)$ be the modal language with knowledge and safe belief:

$$\varphi ::= p \mid \neg\varphi \mid \varphi \wedge \varphi \mid K_i\varphi \mid \Box_i\varphi$$

with $\langle i \cap \geq_i \rangle \varphi \leftrightarrow \neg\Box_i\neg\varphi$. We can now prove that safe belief and knowledge can simulate (conditional) belief:

Fact 3.5. $\mathcal{L}_{DOX} \leq \mathcal{L}_{DOX}(K_i, \langle i \cap \geq_i \rangle)$

Proof. The only non-trivial clause of the translation is the one for conditional belief. We can adapt the translation used in [86, 3.3.6] to our doxastic epistemic setting in a very straightforward way:

$$\tau(B_i^\chi \psi) \ := \ K_i(\tau(\chi) \rightarrow \langle i \cap \geq_i \rangle (\tau(\chi) \wedge \Box_i(\tau(\chi) \rightarrow \tau(\psi))))$$

We now prove that $\mathcal{M}, w \Vdash \varphi$ iff $\mathcal{M}, w \Vdash \tau(\varphi)$. As mentioned the only non-trivial clause is the one for conditional belief. Assume that φ is of the form $B_i^\chi \psi$.

From left to right, assume that $\mathcal{M}, w \Vdash B_i^\chi \psi$. First case: there is no χ-state within $\mathcal{K}_i[w]$. But then $\tau(B_i^\chi \psi) = K_i(\tau(\chi) \rightarrow \langle i \cap \geq_i \rangle (\tau(\chi) \wedge \Box_i(\tau(\chi) \rightarrow \tau(\psi))))$ is trivially true. Second case: we have at least one χ-state within $\mathcal{K}_i[w]$. But then, by truth conditions of $B_i^\chi \psi$, all the minimal χ-states within $\mathcal{K}_i[w]$ are ψ-states (1). Now in any χ-state within $\mathcal{K}_i[w]$ you can $\sim_i \cap \geq_i$-move to one of these minimal states and in such states, we can show that $\Box_i(\tau(\chi) \rightarrow \tau(\psi))$ will hold. For assume there is a state t which is a minimal χ-state within $\mathcal{K}_i[w]$ and $\Box_i(\tau(\chi) \rightarrow \tau(\psi))$ does not hold. Then we must have a state s with $s \leq_i t$, $s \in \mathcal{K}_i[w]$, $\mathcal{M}, s \Vdash \chi$ and $\mathcal{M}, s \nVdash \psi$ (2). But then s is a minimal χ-state within $\mathcal{K}_i[w]$ (3). But (2) and (3) contradicts (1).

From right to left. We prove the contrapositive. Assume that $\mathcal{M}, w \nVdash B_i^\chi \psi$. Then we have a $\leq_i$-minimal v state within $K_i[w] \cap ||\chi||^{\mathcal{M}}$ such that $\mathcal{M}, v \nVdash \psi$

(4). Now assume for a contradiction that $\mathcal{M}, w \Vdash K_i(\tau(\chi) \rightarrow \langle i \cap \geq_i \rangle(\tau(\chi) \wedge \Box_i(\tau(\chi) \rightarrow \tau(\psi)))$. Since $v \in K_i[w]$ and $v \in ||\chi||$ we have $\mathcal{M}, v \Vdash \langle i \cap \geq_i \rangle(\tau(\chi) \wedge \Box_i(\tau(\chi) \rightarrow \tau(\psi)))$. But then we have some state t with $v \sim_i t$ (5), $v \geq_i t$ (6), $\mathcal{M}, t \Vdash \chi$ (7) and $\mathcal{M}, t \Vdash \Box_i(\tau(\chi) \rightarrow \tau(\psi))$ (8). But since v is $\leq_i$-minimal within $K_i[w] \cap ||\chi||^{\mathcal{M}}$, (5), (6) and (7) implies that $t \geq_i v$ (9). But then by (5) and (9), (8) implies that $\mathcal{M}, v \Vdash \tau(\psi)$. But by IH this contradicts (4). QED

On the conceptual side, the preceding fact shows that knowledge, (conditional) belief and safe belief constitutes a natural family of modalities. Moreover considering that the operator $\langle \geq_i \rangle$, scanning the plausibility ordering, was not needed to simulate conditional belief and that it could not be expressed in the doxastic epistemic language (Fact 3.3) indicates that it really belongs to another family of doxastic modalities. The next chapter sheds light on the conceptual relevance of this fact, when discussing definability of concepts such as 'common prior' and the role they play in interactive epistemology, in particular for qualitative agreement theorems. For now let us point out that the natural notion of belief matching $\langle \geq_i \rangle$ is that of *prior* belief $\boxslash_i$. This modality is interpreted by looking at the a priori most plausible elements of the *domain*, rather than the a priori most plausible element of the *information set* of the agent. A notion of conditional *prior* belief can also be defined. Moreover the natural companion modality to these two operators is thus the existential modality $\mathtt{E}$ rather than the knowledge modality. Their semantics follows:

$$\begin{array}{lll} \mathcal{M}, w \Vdash \boxslash_i \varphi & \text{iff} & \text{for all } v \text{ such that } v \in \beta_i(|\mathcal{M}|) \text{ we have } \mathcal{M}, v \Vdash \varphi \\ \mathcal{M}, w \Vdash \boxslash_i^{\psi} \varphi & \text{iff} & \text{for all } v \text{ such that } v \in \beta_i(||\psi||^{\mathcal{M}}) \text{ we have } \mathcal{M}, v \Vdash \varphi \\ \mathcal{M}, w \Vdash \mathtt{E}\varphi & \text{iff} & \text{for some } v \text{ with } v \in |\mathcal{M}| \text{ we have } \mathcal{M}, v \Vdash \varphi \end{array}$$

In the preceding family, $\Box^{\geq_i}$, the dual of $\langle \geq_i \rangle$, can really be reinterpreted as *safe prior* belief, completing the preceding family of doxastic modalities. We will meet it again when discussing definability issues and in the next chapter on interactive epistemology.

Still in the context of interactive epistemology, multi-agent notions of belief and knowledge, such as common belief and common knowledge, will play a crucial role. We leave their introduction to the next chapter as we would like to contrast the notions at work in our logical study of agreement results with respect to epistemic plausibility models with the ones used in the probabilistic approach.

We conclude with two so-called window operators: $[\![\geq_i]\!]$ and $[\![\sim_i]\!]$. The intended meaning of $[\![\geq_i]\!]\varphi$ is that all φ-states are at least as plausible for i as the current one, while $[\![\sim_i]\!]\varphi$ says that all φ-states are considered as epistemically possible given i's current information. Intuitively while 'i knows φ' really means that being a φ-situation is necessary for a state to be considered by i as epistemically possible, $[\![\sim_i]\!]\varphi$ says that being a φ-state is a *sufficient* condition. Operators in this line of thinking have been applied in formal interactive epistemology to the study of the Brandenburger-Keisler [51] paradox. On the technical side, these

operators are central in boolean modal logics and can help to prove completeness for a language with intersection in a very elegant way. Their semantics is:

$$\begin{array}{ll} \mathcal{M}, w \Vdash [\![\geq_i]\!]\varphi & \text{iff for all } v \text{ such that } \mathcal{M}, v \Vdash \varphi \text{ we have } v \leq_i w \\ \mathcal{M}, w \Vdash [\![\sim_i]\!]\varphi & \text{iff for all } v \text{ such that } \mathcal{M}, v \Vdash \varphi \text{ we have } w \sim_i v \\ \mathcal{M}, w \Vdash [\![\geq_i \cap \sim_i]\!]\varphi & \text{iff for all } v \text{ such that } \mathcal{M}, v \Vdash \varphi \text{ we have } w \geq_i v \ \& \ w \sim_i v \end{array}$$

Remark on completeness for doxastic languages. We conclude this section about static doxastic languages with a comment about their axiomatization. Many of the possible sublanguages are just particular cases of canonical multimodal logics and completeness can be obtained through canonical models ([39, ch.4] has details). Such completeness proofs can be extended to prove completeness for such canonical modal logics extended with the existential modality E (see Thm 7.3 in [39, pp.417-418] for a proof). For completeness proofs with a language in which (conditional) belief is a primitive notion, the reader can consult [40, 17].

Extending the language with modalities such as $\langle \geq_i \cap \sim_i \rangle$ bring us into the realm of boolean modal logics. It is possible to axiomatize a language containing $\langle \geq_i \cap \sim_i \rangle$ and the window-type modalities (e.g. $[\![\geq_i \cap \sim_i]\!]$). The crucial axiom is: $[\![\geq_i \cap \sim_i]\!]\varphi \leftrightarrow ([\![\geq_i]\!]\varphi \wedge [\![\sim_i]\!]\varphi)$. See Gargov and Passy [80] for details. Another way to go is to consider a *hybrid* version of our language allowing nominals, i.e. formulas true in exactly one state of the model. In this case intersection can be modally defined on the level of frames. Extending an axiomatization for the basic hybrid language (see [39, 7.3] for details) with $\langle \geq_j \cap \sim_j \rangle i \leftrightarrow (\langle \geq_j \rangle i \wedge \langle \sim_j \rangle i)$ (where i ranges over a countable set of nominals NOM) gives us a complete logic. The idea was first proposed by Gargov et al. [81]. Finally one can consider an even richer hybrid version of $\mathcal{L}_{DOX}(\geq_i, K_i)$ allowing state variables and binders, in which $\langle \geq_i \cap \sim_i \rangle$ become definable at the level of models (see Fact 7.8 for details).

3.2 Dynamic doxastic languages

We have seen a relatively wide range of static doxastic languages. Now given a static doxastic language, it is possible to build up a dynamic doxastic language that matches dynamic belief update. The natural approach is simply to extend the underlying static language with a modality corresponding to each event model of interest, i.e. to every generic soft signal one would like to reason about. For many languages this approach goes very smoothly, so let us start by showing how far we can go in this direction before mentioning some difficulties.

Since this natural approach will work similarly for many doxastic languages, we give the details only for one of them, namely the dynamic language based on $\mathcal{L}_{DOX}(\langle \leq_i \rangle, \langle \geq_i \cap \sim_i \rangle, K_i, \mathtt{E})$. The intrinsic interest of this language is that it matches the structural primitives of epistemic plausibility models, while —

as we have seen in the previous section — being able to express the notion of conditional belief. After giving the semantics of this language, we turn to the issue of completeness via compositional analysis, giving recursion axioms for the previous modalities.

3.2.1 Interpreting dynamic modalities

We define the dynamic doxastic-epistemic language $\mathcal{L}_{DDEL}(\langle \leq_i \rangle \varphi, \langle \geq_i \cap \sim_i \rangle, K_i, \mathtt{E})$ as follows:

$$\varphi ::= p \mid \neg\varphi \mid \varphi \vee \varphi \mid \langle i \rangle \varphi \mid \langle \leq_i \rangle \varphi \mid \langle \geq_i \cap \sim_i \rangle \mid \mathtt{E}\varphi \mid \langle \epsilon, \mathbf{e} \rangle \varphi$$

where i ranges over N, p over a countable set of proposition letters PROP, and $(\epsilon, \mathbf{e})$ ranges over a suitable set of symbols for event models. $\langle \epsilon, \mathbf{e} \rangle \varphi$ means that the event $(\epsilon, \mathbf{e})$ can be executed and after it occurs φ holds. $\langle \leq_i \rangle \varphi$ means that there is a state at most as plausible as the current one where φ holds. $\langle i \rangle \varphi$ means that there is an epistemically possible state where φ holds. Knowledge $K_i\varphi \leftrightarrow \neg\langle i \rangle \neg\varphi$ and the universal modality $\mathtt{A}\varphi \leftrightarrow \neg\mathtt{E}\neg\varphi$ are defined as usual.

All our dynamic doxastic logics will be interpreted on *epistemic plausibility models*, together with *epistemic plausibility event models* and the dynamic operation of *priority update* that takes an epistemic plausibility model and an epistemic plausibility event model as inputs and returns a new epistemic plausibility model. In what follows we often refer to *epistemic plausibility event models* as *event models*.

Semantics. Here is how we interpret the $\mathcal{L}_{DDEL}(\langle \leq_i \rangle, \langle \geq_i \cap \sim_i \rangle, \langle i \rangle, \mathtt{E})$ language. A pointed event model is an event model plus some distinguished element of its domain. To economize on notation we use event symbols in the semantic clause. Also, we write $\mathtt{pre}(e)$ for $\mathtt{pre}_\epsilon(e)$ when things are clear from context. The new clause is of course for dynamic operators $\langle \epsilon, \mathbf{e} \rangle$.

$$\mathcal{M}, w \Vdash \langle \epsilon, \mathbf{e} \rangle \varphi \quad \text{iff} \quad \mathcal{M}, w \Vdash \mathtt{pre}(e) \text{ and } \mathcal{M} \times \epsilon, (w, e) \Vdash \varphi$$

Comment about the preconditions of event models.

The reader might have noticed the presence of the precondition in the clause for the dynamic modality $\langle \epsilon, \mathbf{e} \rangle$. In general one usually takes the precondition language to match the underlying static doxastic language (to our dynamic language). One reason is simply that it saves on the necessity to specify semantics independently for the precondition language. But the most important reason is to be able to carry out a compositional analysis of the dynamic doxastic language (when this is possible) in order to get axiomatic completeness.

3.2.2 Completeness via recursion axioms

Indeed the methodology of dynamic epistemic and doxastic logics revolves around *recursion* axioms. As we have seen for public announcement logic and dynamic epistemic logic in Section 1.6.1, when added on top of some complete static base logic, these axioms fully describe the dynamic component. We recall the well-known *Action-Knowledge* recursion axiom of [20]:

$$[\epsilon, \mathbf{e}]K_i\varphi \leftrightarrow (\mathtt{pre}(e) \rightarrow \bigwedge\{K_i[\epsilon, \mathbf{f}]\varphi \ : \ e \sim_i f\}) \tag{3.1}$$

Let us recall from Section 1.6.1 that together with a complete Hilbert system for the underlying epistemic language and recursion axioms for the booleans, the *Action-Knowledge* recursion axiom is complete for the dynamic epistemic logic of K with respect to epistemic product update on epistemic models. To guarantee that the preceding recursion axiom is finite — and thus that we can define a suitable complexity measure on dynamic formulas by induction on which we can prove soundness of our recursion axioms — the underlying event model has to be finite. See [67, ch.7] for details.

In the same way axiomatizing $\mathcal{L}_{DDEL}(\langle\leq_i\rangle, \langle i \cap \geq_i\rangle, K_i, \mathtt{E})$, or another dynamic doxastic language, can be done by extending complete Hilbert systems for the underlying static language together with recursion axioms for the different modalities (and for booleans and propositional letters). We are able to carry out a complete compositional analysis of the dynamic logic based on some doxastic language $\mathcal{L}$ within $\mathcal{L}$ if we can translate every sentence of the dynamic language using modalities in $\mathcal{L}$ and events modalities $[\epsilon, \mathbf{e}]$ to a sentence of $\mathcal{L}$. (We would say that $\mathcal{L}$ is closed for priority update.) Therefore if we can carry such an analysis and we have a complete axiom system for $\mathcal{L}$, adding the dynamic recursion axioms needed to carry this analysis will give us a complete logic for the corresponding dynamic doxastic logic. It remains to check that the recursion axioms are sound. The axioms for booleans and propositional letters are rather uninteresting and unsurprising. Morever since priority update of an epistemic plausibility model by some event model applies product update to compute the new epistemic relation, the reduction axiom for knowledge is just the familiar *Action-Knowledge*. So let us go directly to the interesting part: recursion axioms for doxastic modalities. We start with the modality $\langle\leq_i\rangle$ scanning the plausibility ordering itself.

Proposition 3.6. *The following law is sound for plausibility change:*

$$\langle\epsilon, \mathbf{e}\rangle\langle\leq_i\rangle\varphi \leftrightarrow (\mathtt{pre}(e) \wedge (\langle\leq_i\rangle \bigvee\{\langle\mathbf{f}\rangle\varphi \ : \ e \simeq_i f\} \vee \mathtt{E}\bigvee\{\langle\mathbf{g}\rangle\varphi \ : \ e <_i g\})) \tag{3.2}$$

Proof. From left to right. Assume that $\mathcal{M}, w \Vdash \langle\epsilon, \mathbf{e}\rangle\langle\leq_i\rangle\varphi$, then we have $\mathcal{M} \times \epsilon, w\mathbf{e} \Vdash \langle\leq_i\rangle\varphi$. This means that we have some vf such that $we \leq_i vf$ (1) and $\mathcal{M} \times \epsilon, v\mathbf{f} \Vdash \varphi$ (2). It follows from (2) that $\mathcal{M}, v \Vdash \langle\epsilon, \mathbf{f}\rangle\varphi$ (3). Moreover it follows from (1) and the definition of priority update that we are in one of two cases.

Case 1: $e <_\epsilon f$ and $v \in |\mathcal{M}|$. But then there is some state v in $|\mathcal{M}|$ and some event f such that $e <_\epsilon f$ and $\mathcal{M} \times \epsilon, v\mathbf{f} \Vdash \varphi$, which gives us $\mathcal{M}, w \Vdash \mathtt{E} \bigvee \{\langle f \rangle \varphi \ : \ e <_i f\}$. **Case 2**: $e \simeq_i^\epsilon f$ (4) and $w \leq_i v$ (5). But then by (3), (4) and (5) we have $\mathcal{M}, w \Vdash \langle \leq_i \rangle \bigvee \{\langle f \rangle \varphi \ : \ e \simeq_i^\epsilon f\}$. The other direction is similar. QED

We now turn to a recursion axiom for the dual of the safe belief modality which together with knowledge can simulate conditional belief modalities.

Proposition 3.7. *The following law is sound for epistemic plausibility change:*

$$\begin{array}{ll} \langle \epsilon, \mathbf{e} \rangle \langle \geq_i \cap\, i \rangle \varphi \leftrightarrow & (\mathtt{pre}(e) \wedge \\ & (\langle i \cap \geq_i \rangle \bigvee \{\langle \mathbf{f} \rangle \varphi \ : \ e \sim_i f \ \& \ e \simeq_i f\} \vee \\ & \langle i \rangle \bigvee \{\langle \mathbf{g} \rangle \varphi \ : \ e \sim_i f \ \& \ e <_i g\})) \end{array}$$

Proof. From left to right. Assume that $\mathcal{M}, w \Vdash \langle \epsilon, \mathbf{e} \rangle \langle \geq_i \cap\, i \rangle \varphi$, then we have $\mathcal{M} \times \epsilon, w\mathbf{e} \Vdash \langle \geq_i \cap\, i \rangle \varphi$. This means that we have some vf such that $we \geq_i vf$ (1) and $we \sim_i vf$ (2) and $\mathcal{M} \times \epsilon, v\mathbf{f} \Vdash \varphi$ (3). It follows from (3) that $\mathcal{M}, v \Vdash \langle \epsilon, \mathbf{f} \rangle \varphi$ (4). By (1) and the definition of priority update we have $e \sim_i f$ (5) and $w \sim_i v$ (6). Moreover it follows from (2) and the definition of priority update that we are in one of two cases. **Case 1**: $e >_\epsilon f$. But then by (6) there exists some state v with $w \sim_i v$ and some event f such that $e <_\epsilon f$ and $\mathcal{M} \times \epsilon, v\mathbf{f} \Vdash \varphi$, which, together with (5), gives us $\mathcal{M}, w \Vdash \langle i \rangle \bigvee \{\langle f \rangle \varphi \ : \ e \sim_i f \ \& \ e <_i f\}$. **Case 2**: $e \simeq_i^\epsilon f$ (7) and $w \geq_i v$ (8). But then by (4), (5), (6), (7), (8) we have $\mathcal{M}, w \Vdash \langle \geq_i \cap i \rangle \bigvee \{\langle f \rangle \varphi \ : \ e \simeq_i^\epsilon f \ \& \ e \sim_i f\}$. The other direction is similar. QED

The recursion axiom for belief follows from the recursion axiom for knowledge, the recursion axiom for safe belief and the translation given in the proof of Fact 3.5. Now the crucial feature of the previous dynamic 'recursion step' is that the order between *action* and *belief* is reversed. This works because, conceptually, the current beliefs already *pre-encode* the beliefs after some specified event. Again we see that, while epistemic recursion axioms reflected agent properties of Perfect Recall and No Miracles [34], doxastic recursion axioms encode 'event-oriented' revision policies, and the same point applies to the principles we will find later in a doxastic temporal setting.

Finally our use of the existential modality reflects our stipulation that strict preference among events can make any two worlds comparable. The recursion axiom for the $\langle \epsilon, \mathbf{e} \rangle \mathtt{E}$ alternation is given by the following

Proposition 3.8. *The following axiom is sound for epistemic plausibility change:*

$$\langle \epsilon, \mathbf{e} \rangle \mathtt{E} \varphi \leftrightarrow (\mathtt{pre}(e) \ \wedge \ (\mathtt{E} \bigvee \{\langle \mathbf{f} \rangle \varphi \ : \ f \in \mathit{Dom}(\epsilon)\})) \tag{3.3}$$

Proof. From left to right. Assume that $\mathcal{M}, w \Vdash \langle \epsilon, \mathbf{e} \rangle \mathtt{E} \varphi$, then we have $\mathcal{M} \times \epsilon, w\mathbf{e} \Vdash \mathtt{E} \varphi$. This means that we have some $vf \in \mathit{Dom}(\mathcal{M} \times \epsilon)$ (1) and $\mathcal{M} \times \epsilon, v\mathbf{f} \Vdash$

φ (2). It follows from (2) that $\mathcal{M}, v \Vdash \langle \epsilon, \mathbf{f} \rangle \varphi$ (3). Moreover it follows from (1) and the definition of priority update that $v \in Dom(\mathcal{M})$ and $f \in Dom(\epsilon)$. But then there exists some state v in $|\mathcal{M}|$ and some event $f \in Dom(\epsilon)$ such that $\mathcal{M} \times \epsilon, v\mathbf{f} \Vdash \varphi$, which gives us $\mathcal{M}, w \Vdash \mathtt{E} \bigvee \{ \langle \mathbf{f} \rangle \varphi \ : \ f \in Dom(\epsilon) \})$ QED

Everything appears to go smoothly but as we mentioned this need not be the case. In the epistemic context, it has been shown that adding public announcement modalities to the epistemic language with common knowledge strictly extends the expressive power of the static language. The same is a fortiori true for dynamic epistemic extensions in general. (See Baltag et al. [20] for a proof.) We say that the language $\mathcal{L}_{EL}(C_G)$ is not closed under product update. It follows that completeness for dynamic epistemic extensions of the epistemic language with common knowledge can no longer be obtained via compositional analysis and must be proven by other means. A completeness proof, based on the filtration argument used by Kozen and Parikh [110]'s in their completeness proof for PDL, is given in [20]. But it was proven that some richer languages are again closed under product update: van Benthem et al. [35] prove this for epistemic PDL and van Benthem and Ikegami [33] for the epistemic μ-calculus (two languages in which common knowledge is expressible). Similar limits of the compositional analysis methodology can occur in the doxastic case and similar solutions are available.

We mentioned local (or unified) plausibility models together with the local priority update rule of Baltag and Smets [17] while discussing representation theorems in terms of doxastic temporal properties. We referred to [17, 40] for corresponding static languages and their axiomatization. The reader who wonders about the recursion axioms (corresponding to local priority update) will find the answer in [17].

This concludes what we wanted to say about dynamic doxastic languages, but the story of logical languages to analyze belief change is not yet complete. Recursion axioms gave the syntactic view of dynamic logics of belief change. Let us now move to the complementary point of view offered by temporal logics.

3.3 Doxastic epistemic temporal languages

The previous section identified the syntactic principles governing the dynamic component of dynamic logics of belief change. But the previous chapter identified the classes of doxastic temporal models generated by stepwise belief revision, so it is now time to discuss how doxastic temporal languages can characterize such frames: another syntactic way of identifying the principles of local, stepwise belief change.

To do so we will consider languages interpreted on *doxastic-epistemic temporal models*, which are simply our old doxastic temporal models extended with

epistemic accessibility relations $\sim_i$. Such languages come in at least two sorts: *branching-time* languages, that are evaluated at both a maximal history and some node on that history (i.e. subsequence of that history); and simple (or linear) *temporal* languages, that are evaluated at a node (finite sequence) in the model. Intuitively the first type of language allows for quantification over possible continuations (histories), while the second type allows only quantification over successors.

3.3.1 Simple doxastic epistemic temporal languages

We start by introducing the doxastic epistemic temporal language $\mathcal{L}_{DET}$ that is the safe belief-free counterpart of the dynamic doxastic language $\mathcal{L}_{DDEL}(\langle\leq_i\rangle, \langle i\cap \geq_i\rangle, \langle i\rangle, \mathtt{E})$ considered in the previous section, extended with one-step backward modalities allowing formulas of the form $\langle e^{-1}\rangle\varphi$, with intuitive meaning 'e has just been executed, and before that φ was true'. The syntax of $\mathcal{L}_{DET}$ is recursively defined as follows:

$$\varphi ::= p \mid \neg\varphi \mid \varphi \vee \varphi \mid \langle e\rangle\varphi \mid \langle e^{-1}\rangle\varphi \mid \langle\leq_i\rangle\varphi \mid \langle i\rangle\varphi \mid \mathtt{E}\varphi,$$

where i ranges over N, e over Σ, and p over proposition letters PROP.

The particular choice of syntax is driven by definability issues. Indeed this language gives the right expressive power to define our earlier doxastic-temporal properties characterizing stepwise belief-revising agents. Other doxastic epistemic temporal languages are very natural. For example a very natural language allows for temporal operators $F\varphi$ and $P\varphi$ with intuitive meaning: 'at some point in the future φ' and at 'some point in the past φ'. This being said, let us fix the semantics of $\mathcal{L}_{DET}$, which we will interpret over nodes h in our trees (cf. [34]):

Definition 3.9 (Truth definition)**.** *Let* $\mathcal{K}_i[h] = \{h' \mid h \sim_i h'\}$

$\mathcal{H}, h \Vdash p$	iff	$h \in V(p)$
$\mathcal{H}, h \Vdash \neg\varphi$	iff	$\mathcal{H}, h \nVdash \varphi$
$\mathcal{H}, h \Vdash \varphi \vee \psi$	iff	$\mathcal{H}, h \Vdash \varphi$ or $\mathcal{H}, h \Vdash \psi$
$\mathcal{H}, h \Vdash \langle e\rangle\varphi$	iff	for some h' with $h' = he$ we have $\mathcal{H}, h' \Vdash \varphi$
$\mathcal{H}, h \Vdash \langle e^{-1}\rangle\varphi$	iff	for some h' with $h'e = h$ we have $\mathcal{H}, h' \Vdash \varphi$
$\mathcal{H}, h \Vdash \langle\leq_i\rangle\varphi$	iff	for some h' with $h \leq_i h'$ we have $\mathcal{H}, h' \Vdash \varphi$
$\mathcal{H}, h \Vdash \langle i\rangle\varphi$	iff	for some h' with $h' \in K_i[h]$ we have $\mathcal{H}, h' \Vdash \varphi$
$\mathcal{H}, h \Vdash \mathtt{E}\varphi$	iff	for some $h' \in H$ we have $\mathcal{H}, h' \Vdash \varphi$

This constituted a typical example of simple (or linear) doxastic epistemic temporal languages. This family of languages forms a natural place in which to reason and axiomatize classes of frames characterizing stepwise belief revising agents. But for some applications such as reasoning about inductive reasoning, i.e. how agents might learn or not learn from inductively given streams of data,

one might need a more expressive family of languages. Let us give an example. Let $h \rightarrow^* h'$ iff there is some finite sequence of events $e^* \in \Sigma^*$ such that $h' = he^*$. Using the future operator $F\varphi$ with semantics:

$$\mathcal{H}, h \Vdash F\varphi \quad \text{iff} \quad \text{for some } h' \text{ with } h \rightarrow^* h' \text{ we have } \mathcal{H}, h' \Vdash \varphi$$

and its dual $G\varphi$, one can for example say that an agent will always know φ ($GK\varphi$) or that after some sequence of events she will know φ ($FK\varphi$) but one cannot say that for every sequence of events, there is point at which she will know φ. Notions involving such an alternation of quantifiers $\forall F$ are precisely what branching-time languages can allow us to express.

3.3.2 Branching-time doxastic temporal languages

As we have seen, some applications call for more complex alternations of quantifiers than the ones simple (or linear) temporal languages allow. We will make use of the expressive power of branching-time doxastic temporal languages as we will bring our logical viewpoint to the study of inductive reasoning in Chapter 5. The particular choice of primitives we are making here is motivated by this application. Let us call this language $\mathcal{L}_{\text{BDET}}$. Its syntax is defined inductively as follows:

$$\varphi := p \mid \neg\varphi \mid \varphi \vee \varphi \mid K_j\varphi \mid B_j\varphi \mid \mathtt{A}\varphi \mid \bigcirc^{-1}\varphi \mid F\varphi \mid P\varphi \mid \forall\varphi$$

where p ranges over a countable set of proposition letters PROP, j over N. $K_j\varphi$ ($B_j\varphi$) reads j knows (believes) that φ. F and P stand for future and past. $\forall\varphi$ means: 'in all continuations φ'. Also: $H\varphi := \neg P\neg\varphi$ and $G\varphi := \neg F\neg\varphi$. Finally $\bigcirc^{-1}\varphi$ means: 'in the previous state φ'.

To be consistent with our main applications of this sort of language in this dissertation we take $\mathcal{L}_{\text{BDET}}$ to be interpreted over an ω-W doxastic epistemic temporal model $\mathcal{H}$, an initial state w, an infinite history $w\epsilon$ and a finite prefix wh of $w\epsilon$ [122, 130].

Definition 3.10. *We give the semantics of $\mathcal{L}_{\text{BDET}}$. We skip the obvious clauses. We take $e \sqsubseteq e'$ to mean that e is an initial segment of e' and let $\mathcal{B}_i[wh] = min_{\leq_i}\mathcal{K}_i[wh]$ be the set of histories that i considers the most plausible at wh.*

$\mathcal{H}, w\epsilon, wh \Vdash p$	iff	$wh \in V(p)$
$\mathcal{H}, w\epsilon, wh \Vdash K_i\varphi$	iff	$\forall vh' \; \forall w\epsilon$ if $vh' \in \mathcal{K}_i[wh] \& vh' \sqsubseteq v\epsilon'$ then $\mathcal{H}, v\epsilon', vh' \Vdash \varphi$
$\mathcal{H}, w\epsilon, wh \Vdash B_i\varphi$	iff	$\forall vh' \; \forall w\epsilon$ if $vh' \in \mathcal{B}_i[wh] \& vh' \sqsubseteq v\epsilon'$ then $\mathcal{H}, v\epsilon', vh' \Vdash \varphi$
$\mathcal{H}, w\epsilon, wh \Vdash \mathtt{A}\varphi$	iff	$\forall vh' \; \forall w\epsilon$ if $vh' \in H \;\&\; vh' \sqsubseteq v\epsilon'$ then $\mathcal{H}, v\epsilon', vh' \Vdash \varphi$
$\mathcal{H}, w\epsilon, wh \Vdash \bigcirc^{-1}\varphi$	iff	$\exists a \in \Sigma \; \exists h' \sqsubseteq \epsilon$ with $h'.a = h$ and $\mathcal{H}, w\epsilon, wh' \Vdash \varphi$
$\mathcal{H}, w\epsilon, wh \Vdash F\varphi$	iff	$\exists e \in \Sigma^* \; \exists h' \sqsubseteq \epsilon$ with $h' = he$ and $\mathcal{H}, w\epsilon, wh' \Vdash \varphi$
$\mathcal{H}, w\epsilon, wh \Vdash P\varphi$	iff	$\exists e \in \Sigma^* \; \exists h' \sqsubseteq \epsilon$ with $h'e = h$ and $\mathcal{H}, w\epsilon, wh' \Vdash \varphi$
$\mathcal{H}, w\epsilon, wh \Vdash \forall\varphi$	iff	$\forall h' \in \mathbb{P}(w)$ s.t. $h \sqsubseteq h$ we have $\mathcal{H}, wh', wh \Vdash \varphi$

Chapter 5 will put branching-time epistemic temporal languages to work in the analysis of inductive learning scenarios; we leave them until then. In the rest of the chapter will we work with simple (i.e. non-branching) doxastic temporal languages.

In the next section we show that simple (non-branching) doxastic epistemic temporal languages give us the right syntax to analyze our earlier structural conditions.

3.3.3 Defining the frame conditions for priority update

Indeed in this section we show that these non-branching doxastic temporal languages give us the right syntax to analyze the structural conditions that emerged earlier in Chapter 2 as the specific properties that characterize agents revising their beliefs with the 'local, stepwise priority updates' of dynamic doxastic logic. We will state semantic *correspondence results* (cf. [39]) for our crucial properties, using somewhat technical axioms in the formal language that simplify the argument. Afterwards, we present some reformulations whose meaning for belief-revising agents is more intuitive.

The key correspondence result

We start with the correspondence result driven by these slightly technical axioms.

Theorem 3.11 (Definability)**.** *Preference Propagation, Preference Revelation and Accommodation are all definable in the doxastic-epistemic temporal language* $\mathcal{L}_{DET}$*.*

- $\mathcal{H}$ *satisfies Preference Propagation iff the following axiom is valid:*

$$\mathtt{E}\langle a\rangle\langle{\leq_i}\rangle\langle b^{-1}\rangle\top \rightarrow ((\langle{\leq_i}\rangle\langle b\rangle p \wedge \langle a\rangle q) \rightarrow \langle a\rangle(q \wedge \langle{\leq_i}\rangle p) \qquad (PP)$$

- $\mathcal{H}$ *satisfies Preference Revelation iff the following axiom is valid:*

$$\mathtt{E}\langle b\rangle\langle{\leq_i}\rangle\langle a^{-1}\rangle\top \rightarrow (\langle a\rangle\langle{\leq_i}\rangle(p \wedge \langle b^{-1}\rangle\top) \rightarrow \langle{\leq_i}\rangle\langle b\rangle p) \qquad (PR)$$

- $\mathcal{H}$ *satisfies Accommodation iff the following axiom is valid:*

$$\begin{aligned}&\mathtt{E}\langle a\rangle\langle{\leq_i}\rangle\langle b^{-1}\rangle\top \\ &\quad\wedge \mathtt{E}\,[\langle a\rangle\,(p_1 \wedge \mathtt{E}\,(p_2 \wedge \langle b^{-1}\rangle\top)\,) \wedge [a]\,(p_1 \rightarrow [{\leq_i}]\neg p_2)] \\ &\qquad\qquad \rightarrow (\,(\langle{\leq_i}\rangle\langle b\rangle q \rightarrow [a]\langle{\leq_i}\rangle q) \\ &\qquad\qquad\quad \wedge (\langle a\rangle\langle{\leq_i}\rangle(r \wedge \langle b^{-1}\rangle\top) \rightarrow \langle{\leq_i}\rangle\langle b\rangle r) \qquad (\mathrm{AC})\end{aligned}$$

Before we prove this correspondence result, let us give some intuitions about the meaning of the previous axioms.

Two intuitive explanations

Here are two ways to grasp the intuitive meaning of our technical axioms.

Reformulation with safe prior belief. We have encountered the *safe belief* modality and its counterpart in the family of 'prior belief', the *safe prior belief* modality $\Box^{\geq}$ (prior beliefs we do not give up under true new information), which is just the universal dual of the existential modality $\langle \geq \rangle$ scanning the converse of $\leq$. Here is how we can then rephrase our earlier axiom:

• $\mathcal{H}$ satisfies Preference Propagation iff the following axiom is valid on $\mathcal{H}$:

$$\mathtt{E}\langle a \rangle \langle \geq_i \rangle \langle b^{-1} \rangle \top \rightarrow (\langle a \rangle \Box^{\geq_i} p \rightarrow \Box^{\geq_i} [b] p) \qquad (PP')$$

A similar reformulation works for Preference Revelation. Principles in this format reverse action modalities and safe belief much like the earlier Knowledge-Action interchange laws. Its intuitive meaning should be understood in terms of acquired safe (prior) beliefs as informative events happen. (PP') states that if at some state after some event a has happened i considers it, a priori, at least as likely that the event b just took place, then if it is possible that after a takes place, i (a priori) safely believes that p, then i already (a priori) safely believes that after b takes place p will hold.

Analogies with recursion axioms One can also understand our formal axioms in their original format with existential modalities by analogy with the dynamic-doxastic recursion axiom for priority update. Here are some cases juxtaposed:

$$\langle \epsilon, \mathbf{e} \rangle \langle \leq_i \rangle p \leftrightarrow (\mathtt{pre}(e) \wedge (\langle \leq_i \rangle \bigvee \{ \langle f \rangle p \ : \ e \simeq_i f \} \vee \mathtt{E} \bigvee \{ \langle g \rangle p \ : \ e <_i g \})) \quad (3.7)$$

$$\mathtt{E}\langle a \rangle \langle \leq_i \rangle \langle b^{-1} \rangle \top \rightarrow (\langle \leq_{\mathtt{i}} \rangle \langle \mathtt{b} \rangle p \rightarrow [\mathtt{a}] \langle \leq_{\mathtt{i}} \rangle p) \qquad (PP)$$

$$\mathtt{E}\langle b \rangle \langle \leq_i \rangle \langle a^{-1} \rangle \top \rightarrow (\langle \mathtt{a} \rangle \langle \leq_{\mathtt{i}} \rangle (p \wedge \langle b^{-1} \rangle \top) \rightarrow \langle \leq_{\mathtt{i}} \rangle \langle \mathtt{b} \rangle p) \qquad (PR)$$

The family resemblance is obvious, and indeed, (PP) and (PR) may be viewed as the two halves of the dynamic-doxastic reduction axiom, transposed to the more general setting of arbitrary doxastic-temporal models.

Let us now get to the proof of the preceding correspondence result.

Proof. We start by proving the case of *Preference Propagation*. We drop agent labels for convenience.

(PP) characterizes Preference Propagation

We first show that (PP) is valid on all models $\mathcal{H}$ based on preference-propagating frames. Assume that $\mathcal{H}, h \Vdash \mathtt{E}\langle a\rangle\langle{\leq_i}\rangle\langle b^{-1}\rangle\top$ (1). Then there are $ja, j'b \in H$ with $ja \leq j'b$ (2). Now let $\mathcal{H}, h \Vdash (\langle{\leq}\rangle\langle b\rangle p \wedge \langle a\rangle q)$ (3). Then there is $h' \in H$ with $h \leq h'$ (4) and $\mathcal{H}, h' \Vdash \langle b\rangle p$ (5), while also $\mathcal{H}, ha \Vdash q$ (6). We must show that $\mathcal{H}, h \vDash \langle a\rangle(q \wedge \langle{\leq_i}\rangle p)$ (7). But, from (2),(4),(6) and *Preference Propagation*, we get $ha \leq h'b$, and we apply the truth definition.

Next, assume that axiom (PP) is valid on a doxastic temporal frame, that is, true under any interpretation of its proposition letters. So, let $ja \leq j'b$ (1), and also $h \leq h'$ (2). Moreover, let $ha, h'b \in H$ (3). First note that (1) automatically verifies the antecedent of (PP) in any node of the tree. Next, we make the antecedent of the second implication in (PP) true at h by interpreting the proposition letter p as just the singleton set of nodes $h'b$, and q as just ha (4). Since (PP) is valid, its consequent will also hold under this particular valuation V. Explicitly we have $\mathcal{H}, V, h \Vdash \langle a\rangle(q \wedge \langle{\leq_i}\rangle p)$. But spelling out what p, q mean there, we get just the desired conclusion that $ha \leq h'b$.

(PR) characterizes Preference Revelation

We start by proving that (PR) is valid on the class of preference-revealing frames. First assume that $\mathcal{H}$ is based on a preference propagating frame (1). Now assume that $\mathcal{H}, h \Vdash \mathtt{E}\langle b\rangle\langle{\leq_i}\rangle\langle a^{-1}\rangle\top$ (2). It follows that there are $jb, j'a \in H$ such that $jb \leq j'a$ (3). Now assume that $\mathcal{H}, h \Vdash \langle a\rangle\langle{\leq_i}\rangle(p \wedge \langle b^{-1}\rangle\top)$ (4). It follows that $ha \in H$ (5) and that $\mathcal{H}, ha \Vdash \langle{\leq_i}\rangle(p \wedge \langle b^{-1}\rangle\top)$ (6). By (6) it follows that there is a history $g \in H$ such that $ha \leq g$ (7) and $\mathcal{H}, g \Vdash (p \wedge \langle b^{-1}\rangle\top)$ (8). In particular $\mathcal{H}, g \Vdash \langle b^{-1}\rangle\top$ (10). From (10) and semantics of $\langle b^{-1}\rangle$ it follows that $g = kb$ (11) for some $k \in H$. From (11) and (8) we have also $\mathcal{H}, kb \Vdash p$ (12). From (12) it follows that $\mathcal{H}, k \Vdash \langle b\rangle p$ (13). We now have to prove that $\mathcal{H}, h \Vdash \langle{\leq_i}\rangle\langle b\rangle p$ (14). By (7) and (11) we have $ha \leq kb$ (15). But by (1),(3),(5),(15) we have by Preference Revelation $h \leq k$ (16). But (14) follows from (16) and (13) by semantics of $\langle{\leq}\rangle$.

Now we assume that $\mathcal{H}$ is not based on a preference-revealing frame (1). We have to find a state and to construct a valuation at which (PR) is not satisfied. It follows from (1) that there are $jb, j'a \in H$ such that $jb \leq j'a$ (2) and that there are $ha, h'b \in H$ such that $ha \leq h'b$ (3) but $h \not\leq h'$ (4). Let us settle $V(p) = \{h'b\}$ (5). By semantics of $\langle b\rangle$ we have $\mathcal{H}, V, h'b \Vdash \langle b^{-1}\rangle\top$ (6). It follows from (5) and (6) that $\mathcal{H}, V, h'b \Vdash p \wedge \langle b^{-1}\rangle\top$ (7). It follows from (3),(7) and the satisfaction condition for $\langle{\leq}\rangle$ that $\mathcal{H}, V, ha \Vdash \langle{\leq}\rangle(p \wedge \langle b^{-1}\rangle\top)$ (8). From (8) and the semantics of $\langle a\rangle$ we have $\mathcal{H}, V, h \Vdash \langle a\rangle\langle{\leq}\rangle(p \wedge \langle b^{-1}\rangle\top)$ (9). It is now sufficient to prove that $\mathcal{H}, V, h \nVdash \langle{\leq_i}\rangle\langle b\rangle p$ (10). But by (5) and semantics of $\langle b\rangle$ it follows that $||\langle b\rangle p|| = \{h'\}$ (11). But (10) follows from (4),(11) and the satisfaction clause of $\langle{\leq}\rangle$.

(AC) **characterizes Accommodation**

Soundness. We start by proving that (AC) is valid on the class of frames satisfying accomodation. First assume that $\mathcal{H}$ is based on an Accommodation-frame (1). Now assume that $\mathcal{H}, g \Vdash \mathtt{E}\langle a\rangle\langle\leq_i\rangle\langle b^{-1}\rangle\top$ (2). It follows that there are $ja, j'b \in H$ such that $ja \leq j'b$ (3). Now assume that $\mathcal{H}, g \Vdash \mathtt{E}\,[\langle a\rangle\,(p_1 \wedge \mathtt{E}\,(p_2 \wedge \langle b^{-1}\rangle\top)\,) \wedge [a]\,(p_1 \rightarrow [\leq_i]\neg p_2)]$ (4). It follows from (4) that there is $h \in H$ such that $\mathcal{H}, h \Vdash \langle a\rangle\,(p_1 \wedge \mathtt{E}\,(p_2 \wedge \langle b^{-1}\rangle\top)\,)$ (5) and $\mathcal{H}, h \Vdash [a]\,(p_1 \rightarrow [\leq_i]\neg p_2)$ (6). From (5) it follows that $\mathcal{H}, ha \Vdash p_1$ (7) and that $\mathcal{H}, ha \Vdash \mathtt{E}\,(p_2 \wedge \langle b^{-1}\rangle\top)$ (8). From (8) it follows that there is some $k \in H$ such that $\mathcal{H}, k \Vdash p_2$ (9) and $\mathcal{H}, k \Vdash \langle b^{-1}\rangle\top)$ (10). From (10) and semantics of $\langle b^{-1}\rangle$ we have $k = h'b$ (11) for some h'. From (6) it follows that $\mathcal{H}, ha \Vdash (p_1 \rightarrow [\leq_i]\neg p_2)$ (12). But (7) and (12) we have $\mathcal{H}, ha \Vdash [\leq_i]\neg p_2$ (13). From (13),(11) and semantics of $[\leq_i]$ it follows that $ha \not\leq h'b$ (14). It follows from (1),(3) and (14) by definition of Accommodation that a and b are accommodating (15).

We first have to prove that $\mathcal{H}, g \Vdash \langle\leq_i\rangle\langle b\rangle q \rightarrow [a]\langle\leq_i\rangle q$ (16a). First assume that $\mathcal{H}, g \Vdash \langle\leq_i\rangle\langle b\rangle q$; it follows that there is a g' such that $g \leq g'$ (17a) and $\mathcal{H}, g' \Vdash \langle b\rangle q$ (18a). It follows from (18a) by semantics of $\langle b\rangle$ that $g'b \in H$ (19a) and $\mathcal{H}, g'b \Vdash q$ (20a). We have to prove that $\mathcal{H}, g \Vdash [a]\langle\leq_i\rangle q$ (21a). Case 1: $ga \notin H$ (22a). But if (22a) then (21a) trivially holds. Case 2: $ga \in H$ (23a). But it follows from (15),(19a),(23a) and (17a) that $ga \leq g'b$ (24a). But (21a) follows from (23a), (24a),(20a) and semantics of $[a]$.

For the other conjunct in the consequent assume we have to prove that $\mathcal{H}, g \Vdash \langle a\rangle\langle\leq_i\rangle(r \wedge \langle b^{-1}\rangle\top) \rightarrow \langle\leq_i\rangle\langle b\rangle r$ (16b). First assume that $\mathcal{H}, g \Vdash \langle a\rangle\langle\leq_i\rangle(r \wedge \langle b^{-1}\rangle\top)$ (17b). It follows from (17b) that $ga \in H$ (18b) and $\mathcal{H}, ga \Vdash \langle\leq_i\rangle(r \wedge \langle b^{-1}\rangle\top)$ (19b). It follows that for some $k \in H$ such that $ga \leq k$ (20b) we have $\mathcal{H}, k \Vdash r$ (21b) and $\mathcal{H}, k \Vdash \langle b^{-1}\rangle\top$ (22b). From (22b) it follows that $k = g'b$ (23b) for some $g' \in H$ (24b). Note that (23b) and (20b) implies that $ga \leq g'b$ (25b). Note also that (23b) and (21b) implies that $\mathcal{H}, g'b \Vdash r$ (26b). But it follows from (15) and (25b) that $g \leq g'$ (27b). Note that (26b) implies by semantics of $\langle b\rangle$ that $\mathcal{H}, g' \Vdash \langle b\rangle r$ (28b). But it follows from (28b) and (26b) that $\mathcal{H}, g \Vdash \langle\leq\rangle\langle b\rangle r$. This concludes the proof for this direction.

Sufficiency. Now we assume that $\mathcal{H}$ is not based on an Accommodation-frame (1). We have to find a state and to construct a valuation at which (AC) is not satisfied. It follows from (1) that there are $ja, j'b \in H$ such that $ja \leq j'b$ (2) and that there are $ha, h'b \in H$ (3) such that $ha \not\leq h'b$ (4) and that a and b are not accommodating (5). Let us settle $V(p_1) = \{ha\}$ (6) and $V(p_2) = \{h'b\}$ (7). We left to the reader to check that $\mathcal{H}, V, j \Vdash \langle a\rangle\langle\leq\rangle\langle b^{-1}\rangle\top$ (8) and that $\mathcal{H}, V, h'b \Vdash (p_2 \wedge \langle b^{-1}\rangle\top)$ (9). By semantics of $[a]$, (6), (7), (3) and semantics of $[\leq]$ it follows that $\mathcal{H}, V, h \Vdash [a]\,(p_1 \rightarrow [\leq_i]\neg p_2)$ (10). From (10), (9) and (6) it is easy to check that $\mathcal{H}, V, h \Vdash \langle a\rangle\,(p_1 \wedge \mathtt{E}\,(p_2 \wedge \langle b^{-1}\rangle\top)\,) \wedge [a]\,(p_1 \rightarrow [\leq_i]\neg p_2)$ (11). From (8),(11) and the semantics of $\mathtt{E}$ it follows that $\mathcal{H}, V, g \Vdash \mathtt{E}\langle a\rangle\langle\leq_i\rangle\langle b^{-1}\rangle\top \wedge \mathtt{E}\,[\langle a\rangle\,(p_1 \wedge \mathtt{E}\,(p_2 \wedge \langle b^{-1}\rangle\top)\,) \wedge [a]\,(p_1 \rightarrow [\leq_i]\neg p_2)]$ (13). There

are two cases in which (5) might hold:

Case 1: $ga, g'b \in H$ (14a), $g \leq g'$ (15a) but $ga \not\leq g'b$ (16a). Let us settle $V(q) = \{g'b\}$ (17a). It follows that $\mathcal{H}, V, g' \Vdash \langle b\rangle q$ (18a). It also follows from (17a), (16a) and semantics of $\langle\leq\rangle$ that $\mathcal{H}, V, ga \nVdash \langle\leq_i\rangle q$ (19a). By semantics of $[a]$ it follows that $\mathcal{H}, V, g \nVdash [a]\langle\leq_i\rangle q$ (20a). But by (15a) and (18a) we have $\mathcal{H}, V, g \Vdash \langle\leq\rangle\langle b\rangle q$ (21a). But (21a) and (19a) implies that $\mathcal{H}, V, g \nVdash \langle\leq_i\rangle\langle b\rangle q \rightarrow [a]\langle\leq_i\rangle q$ (22a). Together with (13), (22a) implies that $\mathcal{H}, V, g \nVdash$ (AC). This concludes the proof for this case.

Case 2: $ga \leq g'b$ (14b) but $g \not\leq g'$ (15b). Let us settle $V(r) = \{g'b\}$ (16b). It is easy to check that $\mathcal{H}, V, g'b \Vdash r \wedge \langle b^{-1}\rangle\top$ (17b). It follows from (17b) by semantics of $\langle\leq\rangle$, (14b) and semantics of $\langle a\rangle$ that $\mathcal{H}, V, g \Vdash \langle a\rangle\langle\leq_i\rangle(r \wedge \langle b^{-1}\rangle\top)$ (18b). It is now sufficient to prove that $\mathcal{H}, V, g \nVdash \langle\leq_i\rangle\langle b\rangle r$ (19b). First note that by (16b) and semantics of $\langle b\rangle$ we have $||\langle b\rangle r|| = \{g'\}$ (20b). But (19b) follows from (15b),(20b) and semantics of $\langle\leq\rangle$. QED

Sahlqvist correspondence

The preceding correspondence proofs can really be seen as *Sahlqvist* substitution arguments. In particular the first two axioms are of a simple form with existential positive antecedents and positive consequents.

Fact 3.12. *The axioms for preference propagation, preference revelation and accomodation can be put in Sahlqvist form.*

Proof. Let us explain this idea (for details see [39, ch.3]). It is well-known that not every modal formula defines a first-order definable class of frames. In fact in the general case, on the level of frames, they really correspond to a fragment of monadic second-order logic. But some of them do correspond to first-order definable classes of frames (**4** characterizes transitive frames). Sahlqvist [141] identifies a large class of syntactically defined modal formulas for which a first-order correspondent can be automatically computed.

In fact (*PP*) and (*PR*) are a particularly simple form of Sahlqvist formula (called very simple Sahlqvist formulas in [39, ch.3]): they are of the form $\varphi \rightarrow \psi$ where φ is built up from propositional letters, $\bot$, $\top$, conjunction and existential modalities, and ψ is positive (i.e. is built up with conjunction, disjunction, existential and universal modalities, one can allow negation but all propositional letters should occur in the scope of an even number of negations). The reader can check that indeed (*PP*) and (*PR*) fall under this scope and hence the correspondences that we found can be justified by a substitution algorithm.

The case of the remaining axiom takes a bit more work. Here is how we can rewrite (AC) into a Sahlqvist form. We split the axiom in two parts and take their conjunction. Indeed the conjunction of two Sahlqvist formulas is still Sahlqvist

(see [39, ch.3]). First let us isolate the antecedent that will be common to both conjuncts.

$$\mathtt{E}\langle a\rangle\langle{\leq_i}\rangle\langle b^{-1}\rangle\top \wedge \mathtt{E}\langle b\rangle p \wedge \langle a\rangle q_2 \qquad \text{(AC-ant)}$$

For the reason mentioned before (AC-ant) is a Sahlqvist antecedent. Now for the first conjunct we define the antecedent as follows:

$$\text{(AC-ant)} \wedge \langle{\leq_i}\rangle(r_1 \wedge \langle b\rangle r_2) \qquad \text{(AC-ant-1)}$$

Again a good candidate to be a Sahlqvist antecedent. The consequent will be the following disjunction of two formulas:

$$\langle a\rangle(q_2 \wedge \langle{\leq_i}\rangle r_2) \vee \mathtt{A}[a]\langle{\leq_i}\rangle p \qquad \text{(AC-consq-1)}$$

This is a positive formula (the main difference with (AC) is that we move a negative formula into the consequent). So (AC-ant-1) $\rightarrow$ (AC-consq-1) is a Sahlqvist formula. We proceed similarly for the second conjunct, simply switching the role of $\langle{\leq_i}\rangle(r_1 \wedge \langle b\rangle r_2)$ and $\langle a\rangle(q_2 \wedge \langle{\leq_i}\rangle r_2)$.

$$\text{(AC-ant)} \wedge \langle a\rangle(q_2 \wedge \langle{\leq_i}\rangle r_2) \qquad \text{(AC-ant-2)}$$

This is still a Sahlqvist antecedent. The consequent is the following disjunction:

$$\langle{\leq_i}\rangle(r_1 \wedge \langle b\rangle r_2) \vee \mathtt{A}[a]\langle{\leq_i}\rangle p \qquad \text{(AC-consq-2)}$$

which is a positive formula. So (AC-ant-2) $\rightarrow$ (AC-consq-2) is a Sahlqvist formula, and hence so is the conjunction: ((AC-ant-1) $\rightarrow$ (AC-consq-1)) $\wedge$ ((AC-ant-2) $\rightarrow$ (AC-consq-2)). QED

3.3.4 A first bit of axiomatics

Section 3.4 will give a completeness result for a particular language, but let us already note that our correspondence analysis is close to explicit formal reasoning about belief-revising doxastic agents. We will give just one illustration, which provides a syntactic counterpart to our earlier Fact 2.19, now suitably stated in our formal language $\mathcal{L}_{DET}$. We use the following bridging axiom to simplify the presentation of the syntactic derivation.

$$\begin{array}{l}\mathtt{E}\,[\langle a\rangle\,(\psi \;\wedge\; \mathtt{E}\,(\varphi \wedge \langle b^{-1}\rangle\top)\,) \;\wedge\; [a]\,(\psi \;\rightarrow [{\leq_i}]\neg\varphi)] \\ \quad\rightarrow\; (\langle a\rangle\langle{\leq_i}\rangle(\chi \wedge \langle b^{-1}\rangle\top) \rightarrow \langle{\leq_i}\rangle\langle b\rangle\chi)\end{array} \qquad (F)$$

First, here is an auxiliary observation:

Fact 3.13. *On total doxastic temporal models the following axiom is valid:*

$$\begin{array}{l}\langle a\rangle(\psi \;\wedge\; \mathtt{E}\,(\varphi \wedge \langle b^{-1}\rangle\top)) \quad\rightarrow \\ \qquad (\,\langle a\rangle(\psi \;\wedge\; \langle{\leq_i}\rangle\varphi) \;\vee\; \mathtt{E}\langle b\rangle(\varphi \wedge \;\langle{\leq_i}\rangle(\psi \wedge \langle a^{-1}\rangle\top)) \quad (\mathit{Tot})\end{array}$$

Now we can state a derivational counterpart to what we had before:

Fact 3.14.

$$\vdash ((PP) \wedge (F)) \rightarrow (\mathrm{AC})$$

$$\vdash ((PR) \wedge (Tot)) \rightarrow (F)$$

Proof. The first theorem is a fact of propositional logic. We focus on the second theorem. First assume that $\langle a\rangle\,(\psi \wedge \mathtt{E}\,(\varphi \wedge \langle b^{-1}\rangle\top)\,) \wedge [a]\,(\psi \rightarrow [\leq_i]\neg\varphi)$ (1). It follows from (1) and (Tot), standard modal reasoning and disjunctive syllogism that $\mathtt{E}\langle b\rangle\langle\leq_i\rangle\langle a^{-1}\rangle\top$ (2). We have thus $\vdash$ (1) $\rightarrow$ (2). Let us call the preceding fact (3). Now assume the antecedent of (F), i.e. $\mathtt{E}\,[\langle a\rangle\,(\psi \wedge \mathtt{E}\,(\varphi \wedge \langle b^{-1}\rangle\top)\,) \wedge [a]\,(\psi \rightarrow [\leq_i]\neg\varphi)]$ (4). It follows from (3), (4) and the fact that **4** holds for $\mathtt{E}$ that $\mathtt{E}\langle b\rangle\langle\leq_i\rangle\langle a^{-1}\rangle\top$ (5). Note that (5) is the antecedent of (PR). It follows from (5) and (PR) that $\langle a\rangle\langle\leq_i\rangle(\chi \wedge \langle b^{-1}\rangle\top) \rightarrow \langle\leq_i\rangle\langle b\rangle\chi$. QED

We also get an immediate counterpart to Fact 2.19:

Corollary 3.15.

$$\vdash ((PP) \wedge (PR) \wedge (Tot)) \rightarrow (\mathrm{AC}) \tag{3.4}$$

But there is more to completeness of temporal languages for stepwise belief revision and we devote almost the rest of the chapter to it. But before we get there let us finish by showing that $\mathcal{L}_{DET}$ was really adequate for our correspondence task and pointing to a few more interesting issues about correspondence theory in the general context of modal logics of belief change.

3.3.5 Variations and extensions of the language

The above doxastic-temporal language is not the only reasonable one. We will mention extensions of this language but we start by a natural question: did we need all the expressive power of $\mathcal{L}_{DET}$?

Weaker languages

Indeed weaker *forward-looking* fragments also make sense, dropping converse and the existential modality. But they do not suffice for our correspondence:

Proposition 3.16 (Undefinability)**.**
Preference Propagation, Preference Revelation and Accommodation are not *definable in the forward-looking fragment of* $\mathcal{L}_{DET}$.

Proof. The reason is the same in all cases: we show that these properties are not preserved under taking *bounded p-morphic images*. Figure 3.3 gives an indication how this works concretely for Preference Propagation. We instantiate the condition for the events a and b. The left-hand model satisfies Preference Propagation *by making its antecedent false*. Note moreover that neither of the models satisfies the consequent of Preference Propagation. Indeed in both cases the preference for the black node over the white node is not propagated by execution of (respectively) a and b. It remains to see that the right-hand model does satisfy the antecedent of Preference Propagation (which follows from the right half of this model) and that the right-hand model is a p-morphic or bounded morphic image of the left-hand model.

Since the 'trick' is that the antecedent falsity is not preserved under taking bounded morphic images and that all three notions share a common antecedent, it is easy to see that a similar argument works for the other two notions. QED

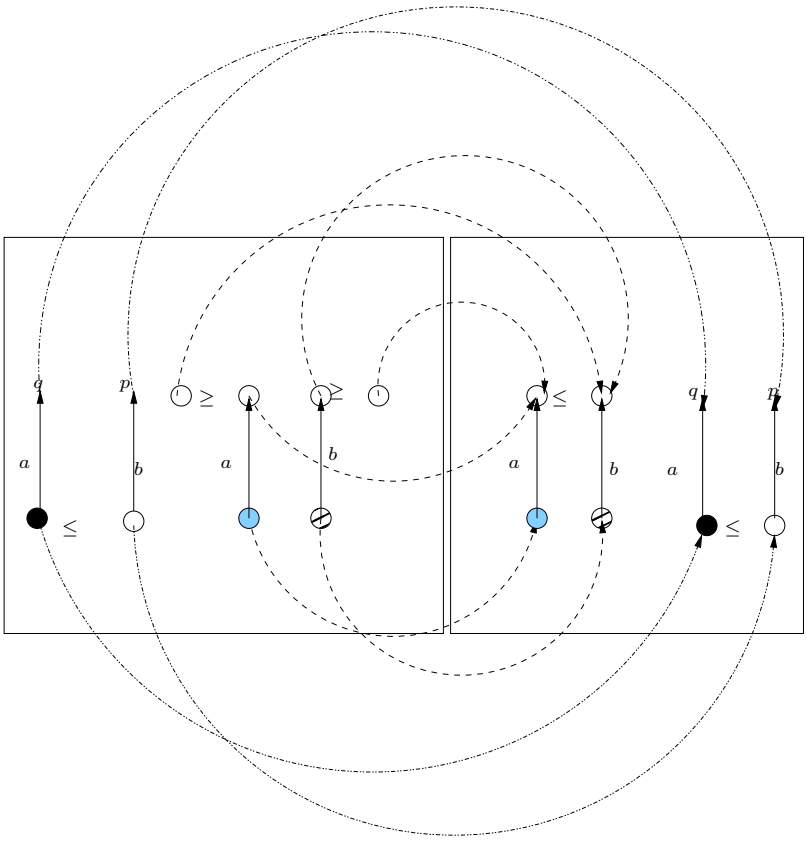

Figure 3.3: Propagation is not preserved under p-morphic images

Richer languages

There is also room for richer languages. For instance, if we want to define the frame property of *synchronicity*, we must introduce an *equilevel relation* in our models, with a corresponding modality for it. While expressing synchronicity then becomes easy, this move is dangerous in principle. van Benthem and Pacuit [34] point at the generally high complexity of tree logics when enriched with this expressive power. Likewise, finer epistemic and doxastic process descriptions require temporal modalities such as "Since" and "Until", beyond the basic operators that we used for matching dynamic doxastic logic directly.

Finally these doxastic temporal languages can also be extended along the doxastic dimension with e.g. operators for beliefs that we have encountered in the section on static doxastic languages.

3.3.6 Further correspondence issues: doxastic and dynamic formulas

In general richer languages uncover new interesting issues for modal correspondence itself. As an example let us consider a formula from the basic doxastic language $\mathcal{L}_{DOX}$.

$$B_i\varphi \rightarrow \varphi \tag{3.5}$$

It has a natural model-theoretic correspondent.

Fact 3.17. *On the class of locally well-founded pointed epistemic plausibility frames, $B_i\varphi \rightarrow \varphi$ characterizes pointed frames in which the pointed (actual) state is in $min_{\leq_i}(\mathcal{K}_i[w])$.*

Proof. For the left to right direction. We prove the contrapositive. Consider a pointed epistemic plausibility frame $\mathcal{F}, w$ such that $w \notin min_{\leq_i}(\mathcal{K}_i[w])$ (1). Then fix the valuation $V(p) = |\mathcal{F}| \setminus \{w\}$ (2).

By local well-foundedness $min_{\leq_i}(\mathcal{K}_i[w])$ is non-empty. Moreover by (1) and (2) we have $min_{\leq_i}(\mathcal{K}_i[w]) \subseteq |\mathcal{F}| \setminus \{w\} = V(p)$. But then $\mathcal{F}, V, w \Vdash B_i p$. But by (2) $\mathcal{F}, V, w \nVdash p$. QED

A natural question arises: in the same way that Sahlqvist formulas constitute a syntactically characterized class of formulas of the basic modal language for which first-order correspondents can be automatically derived, can we define a class of formulas of the basic *doxastic* language for which first-order correspondents can be automatically derived? A similar question has been raised by van Benthem [29, 30] for dynamic epistemic and dynamic doxastic formulas in discussing correspondence results for dynamic languages and extended modal languages in general, respectively.

These questions also apply to other algorithmic approaches to modal correspondence such as structure seeking dialogues as developed by Rahman and Keiff

[136], Keiff [109]. Can structure seeking dialogues for doxastic languages and dynamic languages be defined? How do they relate to dialogical validity games for these logics?

These open questions end our discussion of correspondence theory for modal logics of belief change. We have discussed expressive power, definability and correspondence, but we only made a brief comment on syntactic derivations. The next section balances this by proving axiomatic completeness for a temporal logic of belief revision.

3.4 Axiomatization of protocol-based dynamic logics of belief revision

As the reader might expect the temporal logic of belief revision we will prove axiomatic completeness for, results from the merging of the dynamic doxastic (DDL) and the doxastic temporal (DTL) approaches. Before we lift the curtain on this logic, let us step back briefly to give the background in which it appears. To do so first reconsider public announcement logic (PAL).

PAL is an interesting special case of DEL, in the sense that a public announcement of an epistemic formula φ corresponds to product updating an epistemic model by an epistemic event model based on a singleton with a reflexive accessibility relation and φ as its precondition. The effect is clear from the definition of product update: a public annoucement of φ removes all non-φ states from the domain and preserves the old epistemic relations as far as the new domain allows. We will refer to this extreme way of updating a model on the base of some information as 'hard' update.

We have encountered PAL, DEL (based on product update) and DDL (based on priority update) and the reader might feel that the picture is not complete. The missing element is to be found in the seminal paper van Benthem [29] which explores dynamic logics of belief revision which are a 'soft' counterpart to public announcement. As for public announcement the real input is a formula of some doxastic language, but the effect is very different. One of the operations considered is Lexicographic Upgrade ($\Uparrow$). Lexicographically upgrading a model by a formula φ does not remove states, it changes the plausibility ordering. Precisely it moves all φ states below the non-φ states (thus making the φ-states more plausible than the non-φ-states) and preserves the ordering within these categories. This is an interesting and foundational special case of DDL, just like PAL is an interesting and foundational special case of DEL.

But what about the connection with temporal logics? How does restricting the possible sequences of public announcements by a *protocol* affect the logic of public announcement or the logic of lexicographic upgrade? Indeed as we mentioned in the previous chapter for many applications we would like to focus on the

case in which only some possible streams of information are allowed or possible: communication protocols in interactive epistemology, enumeration protocols in formal learning theory and games as interaction protocols in game theory. The answer in the case of public announcement has been given by Hoshi [106] and [36], which develop the logic TPAL, and prove axiomatic completeness for it, using the following axiomatization:

Axiomatization. The set of formulas of $\mathcal{L}_{TPAL}$ valid on the class of all epistemic temporal models generated from state-dependent public announcement protocols can be axiomatized as follows:

PC	Propostional validities
$(P!\mathbf{K})$	$\vdash [!\varphi](\psi \rightarrow \chi) \rightarrow ([!\varphi]\psi \rightarrow [!\varphi]\chi)$
$(P!p)$	$\vdash \langle!\varphi\rangle p \leftrightarrow (\langle\varphi\rangle\top \wedge p)$
$(P!\neg)$	$\vdash \langle!\varphi\rangle\neg\psi \leftrightarrow (\langle\varphi\rangle\top \wedge \neg\langle!\varphi\rangle\psi)$
$(P!\wedge)$	$\vdash \langle!\varphi\rangle(\psi \wedge \chi) \leftrightarrow (\langle!\varphi\rangle\psi) \wedge \langle!\varphi\rangle\chi)$
$(P!K_i)$	$\vdash \langle!\varphi\rangle K_i\psi \leftrightarrow (\langle\varphi\rangle\top \wedge K_i(\langle!\varphi\rangle\top \rightarrow \langle!\varphi\rangle\psi))$

Table 3.1: Axiom system **TPAL**.

Theorem 3.18 (Completeness of **TPAL**, Hoshi [106], Benthem et al. [36])**.** **TPAL** *is sound and strongly complete with respect to the class of epistemic temporal models generated from a state-dependent public announcement protocol.*

The completeness proof of the preceding result draws on a canonical model construction, indeed no compositional analysis can be carried out when protocols are introduced.

In this section we give the answer in the case of the logic of belief revision, considering what is the effect of protocol-based restrictions on the logic of lexicographic upgrade. To do so we start by introducing protocols that restrict the executable sequences of lexicographic upgrades. Then we choose an underlying doxastic language. Finally we give an axiomatization of protocol-based lexicographic upgrade and prove it complete with respect to the class of doxastic temporal models generated by lexicographically upgrading some doxastic model according to some (belief revision) protocol.

3.4.1 Dynamic logic of protocol-based belief revision

We start with the notion of a state-dependent dynamic belief revision protocol.

Definition 3.19 (State-dependent dynamic belief revision protocol)**.** *Given a doxastic language $\mathcal{L}$ we let $Ptcl(\mathcal{L}) = \{P | P \subseteq \mathcal{L}^*$ and P is closed under initial segments$\}$. Given a doxastic model $\mathcal{M} = \langle W, (\preceq_i)_{i\in N}, V\rangle$, a state-dependent dynamic belief revision protocol is a mapping $\mathtt{p} : W \rightarrow Ptcl(\mathcal{L})$.*

Now how exactly does lexicographic upgrade according to a state-dependent dynamic belief revision protocol generate a temporal forest? A piece of notation: let $h_{(n)}$ stand for the n-th element of h and let $h|n$ stand for the initial segment of h of length n. We will be interested in a particular kind of forests. Namely:

Definition 3.20 (*DoTL* model generated by lexicogaphically upgrading a doxastic model according to a state-dependent dynamic belief revision protocol)**.** *Each initial plausibility model* $\mathcal{M} = \langle W, (\preceq_i)_{i\in N}, V\rangle$ *and each dynamic belief revision protocol* $\mathtt{p} : W \to Ptcl(\mathcal{L})$ *yields a* generated *DoTL* plausibility model $\mathcal{H} = \langle \Sigma, H, (\leq_i)_{i\in N}, \mathbf{V}\rangle$ *as follows:*

- *Let* $\Sigma := \mathcal{L}$.
- *Let* $H_1 := W$ *and for any* $n \geq 1$ *let* $H_{n+1} := \{(w\varphi_1 \dots \varphi_n) | (w\varphi_1 \dots \varphi_{n-1}) \in H_n$ *such that* $\varphi_1 \dots \varphi_n \in \mathtt{p}(w)\}$.

  Finally let $H = \bigcup_{1\leq k} H_k$.
- *If* $h, h' \in H_1$, *then* $h \leq_i h'$ *iff* $h \preceq_i^{\mathcal{M}} h'$.
- *If* $h, h' \in H_1$, *then* $h \equiv h'$.
- *For* $k > 1$ *and for* $h = w\varphi_1 \dots \varphi_{k-1}$, $h' = w'\varphi_1 \dots \varphi_{k-1}$, $h \leq_i h'$ *iff one of the following holds:*
  1. $\mathcal{H}, (w\varphi_1 \dots \varphi_{k-2}) \Vdash \varphi_{k-1}$ *while* $\mathcal{H}, (w'\varphi_1 \dots \varphi_{k-2}) \nVdash \varphi_{k-1}$
  2. $\mathcal{H}, (w\varphi_1 \dots \varphi_{k-2}) \Vdash \varphi_{k-1}$ *iff* $\mathcal{H}, (w'\varphi_1 \dots \varphi_{k-2}) \Vdash \varphi_{k-1}$, *and* $(w\varphi_1 \dots \varphi_{k-2}) \leq_i (w'\varphi_1 \dots \varphi_{k-2})$.
- *For each* $k \geq 1$, *for each* $h, h' \in H_k$, *let* $h \equiv h'$ *iff* $h|(k-1) \equiv h'|(k-1)$ *and* $h_{(k)} = h'_{(k)}$.
- *Let* $wh \in \mathbf{V}(p)$ *iff* $w \in V(p)$.

Now that we have defined the class of doxastic temporal forests generated by lexicographically upgrading some doxastic model according to some state-dependent dynamic belief revision protocol, we can move on to axiomatizing it.

But first of all let us fix the dynamic doxastic language we will prove completeness for. As we have seen lexicographic upgrade only affects the plausibility ordering, so let us first look at a purely doxastic dynamic language ($\mathcal{L}_{DBR}(\mathtt{A}, [\leq_i], \Box^{\geq_i})$) matching our structural primitives. $\Box^{\geq_i}$ is our former safe prior belief operator. $[\leq_i]$ is the modality scanning the plausiblity ordering and $\mathtt{A}$ the universal operator. But the interesting new sentences are of the form $\langle \Uparrow \psi\rangle\varphi$, meaning 'the

protocol allows agents to publicly receive the soft information that ψ, and after they receive this information φ holds'. ($\mathcal{L}_{DBR}(\mathtt{A}, [\leq_i], \Box^{\geq_i})$) is defined by the following inductive syntax

$$\varphi ::= p \mid \neg\varphi \mid \varphi \vee \varphi \mid \langle\leq_i\rangle\varphi \mid \langle\geq_i\rangle\varphi \mid \mathtt{A}_{\equiv}\varphi \mid \langle\Uparrow \psi\rangle\varphi,$$

where i ranges over N, ψ over ($\mathcal{L}_{DOX}(\mathtt{E}, [\leq_i], \Box^{\geq_i})$) (the underlying static doxastic language) and p over a countable set of proposition letters PROP. We make use of the usual shortcuts.

This language ($\mathcal{L}_{DBR}(\mathtt{A}, [\leq_i], \Box^{\geq_i})$, henceforth $\mathcal{L}_{DBR}$) will be interpreted over nodes wh in our trees, where w is a sequence of length 1 and h is possibly the empty sequence, as follows:

$\mathcal{H}, wh \Vdash p$	iff	$wh \in V(p)$
$\mathcal{H}, wh \Vdash \neg\varphi$	iff	$\mathcal{H}, wh \nVdash \varphi$
$\mathcal{H}, wh \Vdash \varphi \vee \psi$	iff	$\mathcal{H}, wh \Vdash \varphi$ or $\mathcal{H}, wh \Vdash \psi$
$\mathcal{H}, wh \Vdash \langle\leq_i\rangle\varphi$	iff	for some h' with $wh \leq_i h'$ we have $\mathcal{H}, h' \Vdash \varphi$
$\mathcal{H}, wh \Vdash \langle\geq_i\rangle\varphi$	iff	for some h' with $h' \leq_i wh$ we have $\mathcal{H}, h' \Vdash \varphi$
$\mathcal{H}, wh \Vdash \mathtt{A}_{\equiv}\varphi$	iff	for all h' such that $h' \equiv wh$ we have $\mathcal{H}, h' \Vdash \varphi$
$\mathcal{H}, wh \Vdash \langle\Uparrow \varphi\rangle\psi$	iff	for some $h' \in H$ with $h' = wh\varphi$ we have $\mathcal{H}, h' \Vdash \psi$

Remember from 3.1 that prior conditional belief will be definable in this language as follows:

$$\boxslash_i^{\psi}\varphi \;\leftrightarrow\; \mathtt{A}(\psi \rightarrow \langle\geq_i\rangle(\psi \wedge \Box^{\geq_i}(\psi \rightarrow \varphi)))$$

Thus it really comes for free in the previous language, leaving us able to focus on the axiomatization of the lighter language ($\mathcal{L}_{DBR}$).

3.4.2 Proving axiomatic completeness.

The axiomatization **DOX** of the static part is well-understood (see Table 3.2 for its details). Together with **DOX**, the axiom system $\langle\Uparrow\rangle$**DOX** for the 'dynamic-temporal' component – we are now really at the interface of both approaches — given in Table 3.3 is a complete axiomatization of the validities of $\mathcal{L}_{DBR}$ over the class of doxastic temporal forests generated by the lexicographic upgrade of a doxastic plausibility model according to a state-dependent dynamic belief revision protocol. Let us now move on to the completeness proof.

Completeness

Let us fix a basic definition before explaining the strategy of the proof.

Definition 3.21 ($\langle\Uparrow\rangle$**DOX**-MCS)**.** *A set of formulas Γ is a $\langle\Uparrow\rangle$**DOX**-maximally consistent set (henceforth a $\langle\Uparrow\rangle$**DOX**-MCS) if Γ is $\langle\Uparrow\rangle$**DOX**-consistent, and any set of formulas properly containing Γ is $\langle\Uparrow\rangle$**DOX**-inconsistent.*

$(\mathbf{K}[\leq_i])$	$\vdash (\varphi \to \psi) \to ([\leq_i]\varphi \to [\leq_i]\psi)$
$(\mathbf{T}[\leq_i])$	$\vdash [\leq_i]\varphi \to \varphi$
$(\mathbf{4}[\leq_i])$	$\vdash [\leq_i]\varphi \to [\leq_i][\leq_i]\varphi$
$(\mathbf{K}K_i)$	$\vdash (\varphi \to \psi) \to (K_i\varphi \to K_i\psi)$
$(\mathbf{T}K_i)$	$\vdash K_i\varphi \to \varphi$
$(\mathbf{5}K_i)$	$\vdash \langle i\rangle\varphi \to K_i\langle i\rangle\varphi$
$(\mathbf{K}\mathtt{A}_{\equiv})$	$\vdash (\varphi \to \psi) \to (\mathtt{A}_{\equiv}\varphi \to \mathtt{A}_{\equiv}\psi)$
$(\mathbf{T}\mathtt{A}_{\equiv})$	$\vdash \mathtt{A}_{\equiv}\varphi \to \varphi$
$(\mathbf{5}\mathtt{A}_{\equiv})$	$\vdash \mathtt{E}_{\equiv}\varphi \to \mathtt{A}_{\equiv}\mathtt{E}_{\equiv}\varphi$
$(\leq - \geq)$	$\vdash (\langle\leq_i\rangle\Box^{\geq_i}\varphi \to \varphi) \wedge (\langle\geq_i\rangle[\leq_i]\varphi \to \varphi)$
$(\leq \subseteq \equiv)$	$\vdash \mathtt{A}_{\equiv}\varphi \to [\leq_i]\varphi$
$(\sim \subseteq \equiv)$	$\vdash \mathtt{A}_{\equiv}\varphi \to K_i\varphi$

Table 3.2: Axiomatization of the static part — **DOX**.

$(\mathbf{K}\langle\Uparrow\rangle)$	$\vdash (\varphi \to \psi) \to (\langle\Uparrow \chi\rangle\varphi \to \langle\Uparrow \chi\rangle\psi)$
$(\langle\Uparrow\rangle\text{PROP})$	$\vdash \langle\Uparrow \varphi\rangle p \leftrightarrow \langle\Uparrow \varphi\rangle\top \wedge p$, for all $p \in$ PROP.
$(\langle\Uparrow\rangle\neg)$	$\vdash \langle\Uparrow \varphi\rangle\neg\psi \leftrightarrow \langle\Uparrow \varphi\rangle\top \wedge \neg\langle\Uparrow \varphi\rangle\psi$
$(\langle\Uparrow\rangle\wedge)$	$\vdash \langle\Uparrow \varphi\rangle(\psi \wedge \chi) \leftrightarrow \langle\Uparrow \varphi\rangle\psi \wedge \langle\Uparrow \varphi\rangle\chi$
$(\langle\Uparrow\rangle[\leq_i])$	$\vdash \langle\Uparrow \varphi\rangle[\leq_i]\psi \leftrightarrow \langle\Uparrow \varphi\rangle\top \wedge$ $[(\varphi \to \mathtt{A}_{\equiv}(\neg\varphi \to \neg\langle\Uparrow \varphi\rangle\neg\psi))$ $\wedge(\varphi \to [\leq_i]\neg\langle\Uparrow \varphi\rangle\neg\psi)$ $\wedge[\leq_i](\neg\varphi \to \neg\langle\Uparrow \varphi\rangle\neg\psi)]$
$(\langle\Uparrow\rangle K_i)$	$\vdash \langle\Uparrow \varphi\rangle K_i\psi \leftrightarrow \langle\Uparrow \varphi\rangle\top \wedge K_i(\langle\Uparrow \varphi\rangle\top \to \langle\Uparrow \varphi\rangle\psi)$
$(\langle\Uparrow\rangle[\mathtt{A}_{\equiv}])$	$\vdash \langle\Uparrow \varphi\rangle\mathtt{A}_{\equiv}\psi \leftrightarrow \langle\Uparrow \varphi\rangle\top \wedge \mathtt{A}_{\equiv}(\langle\Uparrow \varphi\rangle\top \to \langle\Uparrow \varphi\rangle\psi)$

Table 3.3: Axiomatization of the dynamic part.

The strategy of the proof is as follows. We start by defining a $\langle\Uparrow\rangle$**DOX**-canonical initial epistemic plausibility model and then show how it can be unfolded into a canonical doxastic epistemic forest. We use a labelling function λ that associates with each history (finite sequence) in the canonical forest a set of formulas. Lemma 3.24 proves that it associates a $\langle\Uparrow\rangle$**DOX**-MCS with each history. We then prove a truth lemma (Lemma 3.25) with the main induction on the complexity of the formulas and a sub-induction on the length of the history considered. Finally we prove that any subforest of the canonical forest generated by an initial state is isomorphic to the forest generated by lexicographically upgrading some epistemic plausibility model according to some state-dependent belief revision protocol. Completeness follows.

Definition 3.22 (Canonical initial epistemic plausibility model)**.**

We define the $\langle\Uparrow\rangle$**DOX** *canonical initial epistemic plausibility model* $\mathcal{M}_0^{\Sigma} =$

$\langle W^0, (\leq_i^0)_{i\in N}, (\geq_i^0)_{i\in N}, (\sim_i^0)_{i\in N}, \equiv^0, V^0\rangle$ *as follows:*

- W^0 *is the set of* $\langle\Uparrow\rangle$**DOX**-*MCSs*
- *For each* $w, v \in W^0$, *define* $w \leq_i^0 v$ *iff* $\{\varphi \mid [\leq_i]\varphi \in w\} \subseteq v$
- *For each* $w, v \in W^0$, *define* $w \geq_i^0 v$ *iff* $\{\varphi \mid \Box^{\geq_i}\varphi \in w\} \subseteq v$
- *For each* $w, v \in W^0$, *define* $w \sim_i^0 v$ *iff* $\{\varphi \mid K_i\varphi \in w\} \subseteq v$
- *For each* $w, v \in W^0$, *define* $w \equiv^0 v$ *iff* $\{\varphi \mid \mathtt{A}_{\equiv}\varphi \in w\} \subseteq v$
- *Finally define* $V^0(p) = \{w \in H^0 \mid p \in w\}$.

We can now define a canonical doxastic forest by unfolding the canonical initial model. In the construction a labeling function λ associates with each history (finite sequence) a set of formulas.

Definition 3.23 (Canonical doxastic epistemic forest)**.**
$\langle\Uparrow\rangle$**DOX** *canonical forest* $\mathcal{H}^\Sigma = \langle H^\Sigma, \lambda, \leq_i^\Sigma, \geq_i^\Sigma, \sim_i^\Sigma, V^\Sigma\rangle$ *is defined as follows:*

- $H^0 = W^0$
- *For each* $h \in H^0$ *let* $\lambda(h) = h$
- $H_{n+1} = \{h\varphi \mid h \in H_n \text{ and } \langle\Uparrow \varphi\rangle\top \in \lambda(h)\}$
- *For each* $k > 0$ *and* $h\varphi \in H_k$ *let* $\lambda(h\varphi) = \{\psi \mid \langle\Uparrow \varphi\rangle\psi \in \lambda(h)\}$
- $H^\Sigma = \bigcup_{k\geq 0} H^k$
- *For each* $h, h' \in H^0$, *define* $h \leq_i^\Sigma h'$ *iff* $\{\varphi \mid [\leq_i]\varphi \in \lambda(h)\} \subseteq \lambda(h')$
- *For each* $h, h' \in H^0$, *define* $h \geq_i^\Sigma h'$ *iff* $\{\varphi \mid \Box^{\geq_i}\varphi \in \lambda(h)\} \subseteq \lambda(h')$
- *For each* $h, h' \in H^0$, *define* $h \sim_i^\Sigma h'$ *iff* $\{\varphi \mid K_i\varphi \in \lambda(h)\} \subseteq \lambda(h')$
- *For each* $h, h' \in H^0$, *define* $h \equiv^\Sigma h'$ *iff* $\{\varphi \mid \mathtt{A}_{\equiv}\varphi \in \lambda(h)\} \subseteq \lambda(h')$
- *For each* $k > 0$ *and* $h\varphi, h'\psi \in H^k$, *define* $h\varphi \leq^\Sigma h'\psi$ *iff* $h = h'$, $\varphi = \psi$ *and one of the following holds:*
  1. $\varphi \in \lambda(h)$ *while* $\varphi \notin \lambda(h')$
  2. $\varphi \in \lambda(h)$ *iff* $\varphi \in \lambda(h')$, *and* $h \leq_i h'$.
- *For each* $k > 0$ *and* $h\varphi, h'\psi \in H^k$, *define* $h\varphi \geq^\Sigma h'\psi$ *iff* $h'\psi \leq^\Sigma h\varphi$

- *For each $k > 0$ and for each $h\varphi, h'\psi \in H^k$, define $h\varphi \equiv^\Sigma h'\psi$ iff $h \equiv^\Sigma h'$ and $\varphi = \psi$.*
- *For each $p \in$ PROP, define $V^\sigma(p) = \{h \in H^\Sigma \mid p \in \lambda(h_{(1)})\}$.*

We will now prove that the labelling function λ associates a $\langle\Uparrow\rangle$**DOX**-MCS with each history (finite sequence).

Lemma 3.24. *For each $k \geq 0$, for each $h \in H^k$, $\lambda(h)$ is a $\langle\Uparrow\rangle$**DOX**-MCS.*

Proof. The proof is by induction on k. The base case holds by definition. Assume that the claim holds for $k = n$. Now assume that $h\varphi \in H^{n+1}$. By IH $\lambda(h)$ is a MCSs. Moreover by construction we have $\langle\Uparrow \varphi\rangle\top \in \lambda(h)$ (1). Let $\varphi \in \mathcal{L}_{DBR}$. Since $\lambda(h)$ is a MCSs we have either $\langle\Uparrow \varphi\rangle\psi \in \lambda(h)$ or $\neg\langle\Uparrow \varphi\rangle\psi \in \lambda(h)$. If $\langle\Uparrow \varphi\rangle\psi \in \lambda(h)$, then by construction $\psi \in \lambda(h\varphi)$. If instead $\neg\langle\Uparrow \varphi\rangle\psi \in \lambda(h)$ then by (1) and $(\langle\Uparrow\rangle\neg)$ we have $\langle\Uparrow \varphi\rangle\neg\psi \in \lambda(h)$. It follows by construction that $\neg\psi \in \lambda(h\varphi)$. Therefore for each $\varphi \in \mathcal{L}_{DBR}$ we have either $\psi \in \lambda(h\varphi)$ or $\neg\psi \in \lambda(h\varphi)$.

Now we have to prove that $\lambda(h\varphi)$ is consistent. Assume for a contradiction that it is not. Then by definition we have a finite set of formulas $\{\varphi_1, \ldots, \varphi_m\} \subseteq \lambda(h\varphi)$ such that $\vdash (\bigwedge_{i=1}^m \varphi_i) \rightarrow \bot$ (2). It follows from (2) by standard modal reasoning that $\vdash \langle\Uparrow \varphi\rangle\top \rightarrow \langle\Uparrow \varphi\rangle(\bigvee_{i=1}^m \neg\varphi_i)$. It follows by $(\mathbf{K}\langle\Uparrow\rangle)$ again that $\vdash \langle\Uparrow \varphi\rangle\top \rightarrow (\bigvee_{i=1}^m \langle\Uparrow \varphi\rangle\neg\varphi_i)$ (3). By (1) and (3) it follows that $(\bigvee_{i=1}^m \langle\Uparrow \varphi\rangle\neg\varphi_i) \in \lambda(h)$ (4). But since by IH $\lambda(h)$ is a MCSs, there is some j such that $1 \leq j \leq m$ and $\langle\Uparrow \varphi\rangle\neg\varphi_j \in \lambda(h)$ (5). From (5) and $(\langle\Uparrow\rangle\neg)$ we have $\neg\langle\Uparrow \varphi\rangle\varphi_j \in \lambda(h)$ (6). From (2) it follows by contruction that $\langle\Uparrow \varphi\rangle\varphi_i \in \lambda h$ for each i such that $1 \leq i \leq m$ (7). But (6) and (7) together contradicts the fact that $\lambda(h)$ is consistent. It follows by reduction that $\lambda(h\varphi)$ is consistent. QED

We are now ready to prove a Truth Lemma.

Lemma 3.25 (Truth Lemma)**.**

For every $\varphi \in \mathcal{L}_{DBR}$, for each $h \in H^\Sigma$ we have:

$$\varphi \in \lambda(h) \text{ iff } \mathcal{H}^\Sigma, h \Vdash \varphi$$

Proof. The proof is by induction on the complexity of φ. Base case (for atomic formulas) and boolean cases are easy.

[$\mathtt{A}_\equiv$**-modality.**] *From left to right.* Assume that $\mathtt{A}_\equiv\psi \in \lambda(h)$ (0). There are two cases. Either $h \in H^0$ (1) or $h \in (H^\Sigma - H^0)$ (2).

Let us consider the first case. Assume that $h, h' \in H^0$ and that $h \equiv^\Sigma h'$ (3). It follows from (3) by construction that $\{\varphi \mid \mathtt{A}_\equiv\varphi \in \lambda(h)\} \subseteq \lambda(h')$ (4). From (4) and (0) we know in particular that $\psi \in \lambda(h')$ (5). By (5) and the IH of the main induction on formulas it follows that $\mathcal{H}^\Sigma, h' \Vdash \psi$ (6). Since h' was arbitrary, it follows therefore from (6) and the truth definition of $\mathtt{A}_\equiv$ that $\mathcal{H}^\Sigma, h \Vdash \mathtt{A}_\equiv\psi$ (7).

Let us now consider the second case: $h \in (H^{\Sigma} - H^0)$ (2). Without loss of generality let us assume that h is of the form $w\varphi_1 \dots \varphi_{n+1}$ (8). From (8) and (0) it follows by construction that $\langle \Uparrow \varphi_{n+1}\rangle \mathtt{A}_{\equiv}\psi \in \lambda(w\varphi_1 \dots \varphi_n)$ (9). Since by Lemma 3.24 $\lambda(w\varphi_1 \dots \varphi_n)$ is a $\langle \Uparrow \rangle$**DOX**-MCS, it follows from (9) and ($\langle \Uparrow \rangle[\mathtt{A}_{\equiv}]$) that $\langle \Uparrow \varphi_{n+1}\rangle \top \in \lambda(w\varphi_1 \dots \varphi_n)$ (10) and $\mathtt{A}_{\equiv}(\langle \Uparrow \varphi_{n+1}\rangle \top \rightarrow \langle \Uparrow \varphi_{n+1}\rangle \psi) \in \lambda(w\varphi_1 \dots \varphi_n)$ (11). Iterating the same argument we find that:

$$\begin{aligned} \mathtt{A}_{\equiv}(\langle \Uparrow \varphi_1\rangle \top \rightarrow (\langle \Uparrow \varphi_2\rangle \top \rightarrow \dots \\ (\langle \Uparrow \varphi_n\rangle \top \rightarrow (\langle \Uparrow \varphi_{n+1}\rangle \top \rightarrow \\ \langle \Uparrow \varphi_{n+1}\rangle \psi)) \dots)) \in \lambda(w) \end{aligned} \tag{12}$$

and that for each i, $1 \leq k \leq n$

$$\langle \Uparrow \varphi_{k+1}\rangle \top \in \lambda(w\varphi_1 \dots \varphi_k) \tag{13}$$

and that

$$\langle \Uparrow \varphi_1\rangle \top \in \lambda(w) \tag{14}$$

Now assume that $h \equiv^{\Sigma} h'$ (15). It follows from (15) by construction and some easy induction that h' is of the form $v\varphi_1 \dots \varphi_{n+1}$ (16). Similarly it follows from (15) by construction and some induction that $w \equiv^{\Sigma} v$ (17). It follows by construction from (17) and (12) that $(\langle \Uparrow \varphi_1\rangle \top \rightarrow (\langle \Uparrow \varphi_2\rangle \top \rightarrow \dots (\langle \Uparrow \varphi_n\rangle \top \rightarrow (\langle \Uparrow \varphi_{n+1}\rangle \top \rightarrow \langle \Uparrow \varphi_{n+1}\rangle \psi)) \dots)) \in \lambda(v)$ (18). But it is then easy to check that from (18), (13) and (14) we have $\psi \in h'$ (19). It follows from (19) by the main IH that $\mathcal{H}, h' \Vdash \psi$. But since h' was arbitrary, we have: $\mathcal{H}^{\Sigma}, h \Vdash \mathtt{A}_{\equiv}\psi$.

[$\mathtt{A}_{\equiv}$**-modality.**] *From right to left.* Assume that $\mathtt{A}_{\equiv}\psi \notin \lambda(h)$ (0). There are two cases. Either $h \in H^0$ (1) or $h \in (H^{\Sigma} - H^0)$ (2).

Let us consider the first case. Assume that $\mathtt{A}_{\equiv}\psi \notin \lambda(h)$ (0). We have to prove that $\mathcal{H}^{\Sigma}, h \nVdash \mathtt{A}_{\equiv}\psi$. To prove (0) it is sufficient to find a MCSs $\lambda(h')$ such that $\psi \notin \lambda(h')$, but $\{\varphi \mid \mathtt{A}_{\equiv}\varphi \in \lambda(h)\} \subseteq \lambda(h')$. By the Lindenbaum Lemma, it is sufficient to show that $v_0 = \{\neg\psi\} \cup \{\varphi \mid \mathtt{A}_{\equiv}\varphi \in \lambda(h)\}$ is consistent. Assume for a contradiction that it is not. Then we have a finite set of formulas $\varphi_1 \dots \varphi_m \in \{\varphi \mid \mathtt{A}_{\equiv}\varphi \in \lambda(h)\}$ such that $(\bigwedge_{i=1}^{m} \varphi_i) \rightarrow \psi$ (3). But then it follows by standard modal reasoning that $(\bigwedge_{i=1}^{m} \mathtt{A}_{\equiv}\varphi_i) \rightarrow \mathtt{A}_{\equiv}\psi$ (4). But since $\lambda(h)$ is a MCS, it follows from (4) that $\mathtt{A}_{\equiv}\psi \in \lambda(h)$ which contradicts (0).

Let us now consider the other case. Assume WLOG that $h = w\varphi_1 \dots \varphi_{n+1}$. Now assume that $\mathtt{A}_{\equiv}\psi \notin \lambda(h)$ (0). It follows from (0) by maximality of $\lambda(h)$ and construction that $\langle \Uparrow \varphi_{n+1}\rangle \neg \mathtt{A}_{\equiv}\psi \in \lambda(w\varphi_1 \dots \varphi_n)$ (1). It follows from (1) by ($\langle \Uparrow \rangle\neg$) that $\langle \Uparrow \varphi_{n+1}\rangle \top \in \lambda(w\varphi_1 \dots \varphi_n)$ (2) and $\neg\langle \Uparrow \varphi_{n+1}\rangle \mathtt{A}_{\equiv}\psi \in \lambda(w\varphi_1 \dots \varphi_n)$ (3). From (2) and (3), an easy argument gives us by ($\langle \Uparrow \rangle[\mathtt{A}_{\equiv}]$) that $\neg\mathtt{A}_{\equiv}(\langle \Uparrow \varphi_{n+1}\rangle \top \rightarrow \langle \Uparrow \varphi_{n+1}\rangle \psi) \in \lambda(w\varphi_1 \dots \varphi_n)$ (4). Repeating this argument we find that $\neg\mathtt{A}_{\equiv}(\langle \Uparrow \varphi_1\rangle \top \rightarrow \langle \Uparrow \varphi_1\rangle(\langle \Uparrow \varphi_2\rangle \top \rightarrow (\dots \langle \Uparrow \varphi_n\rangle(\langle \Uparrow \varphi_{n+1}\rangle \top \rightarrow \langle \Uparrow \varphi_{n+1}\rangle \psi) \dots))) \in \lambda(w)$ (5).

We will now prove that there is a v such that $w\varphi_1 \dots \varphi_{n+1} \equiv v\varphi_1 \dots \varphi_{n+1}$ and $\psi \notin \lambda(v\varphi_1 \dots \varphi_{n+1})$ (7). First take the following set $v_0 = \{\chi \mid A_{\equiv}\chi \in \lambda(w)\} \cup \{\neg\langle\Uparrow \varphi_1\rangle\top \rightarrow \langle\Uparrow \varphi_1\rangle(\langle\Uparrow \varphi_2\rangle\top \rightarrow (\dots\langle\Uparrow \varphi_n\rangle(\langle\Uparrow \varphi_{n+1}\rangle\top \rightarrow \langle\Uparrow \varphi_{n+1}\rangle\psi)\dots))\}$. Assume for a contradiction that v_0 is inconsistent (8). It follows from (8) that there is a finite set of formulas $\{\chi_1, \dots \chi_m\} \subseteq \{\chi \mid A_{\equiv}\chi \in \lambda(w)\}$ such that $\vdash (\bigwedge_{i=1}^{m} \chi_i) \rightarrow (\langle\Uparrow \varphi_1\rangle\top \rightarrow \langle\Uparrow \varphi_1\rangle(\langle\Uparrow \varphi_2\rangle\top \rightarrow (\dots\langle\Uparrow \varphi_n\rangle(\langle\Uparrow \varphi_{n+1}\rangle\top \rightarrow \langle\Uparrow \varphi_{n+1}\rangle\psi)\dots)))$ (9). But from (9) and standard modal reasoning we find that $\vdash (\bigwedge_{i=1}^{m} \mathtt{A}_{\equiv}\chi_i) \rightarrow \mathtt{A}_{\equiv}(\langle\Uparrow \varphi_1\rangle\top \rightarrow \langle\Uparrow \varphi_1\rangle(\langle\Uparrow \varphi_2\rangle\top \rightarrow (\dots\langle\Uparrow \varphi_n\rangle(\langle\Uparrow \varphi_{n+1}\rangle\top \rightarrow \langle\Uparrow \varphi_{n+1}\rangle\psi)\dots)))$ (10). But since $\lambda(h)$ is a MCS, it follows from (10) that $\mathtt{A}_{\equiv}(\langle\Uparrow \varphi_1\rangle\top \rightarrow \langle\Uparrow \varphi_1\rangle(\langle\Uparrow \varphi_2\rangle\top \rightarrow (\dots\langle\Uparrow \varphi_n\rangle(\langle\Uparrow \varphi_{n+1}\rangle\top \rightarrow \langle\Uparrow \varphi_{n+1}\rangle\psi)\dots))) \in \lambda(h)$ which contradicts (5). Thus by reduction v_0 is consistent. By the Lindenbaum Lemma we can extend v_0 to a maximally consistent v. But by construction $v \in H^{\Sigma}$ (11). Since $\{\chi \mid A \equiv \chi \in \lambda(w)\} \subseteq v_0 \subseteq v$ it follows by construction that $w \equiv v$. But then an easy induction shows that for every $1 \leq j \leq n+1$ we have:

$$w\varphi_1 \dots \varphi_j \equiv v\varphi_1 \dots \varphi_j \qquad (A_{\equiv})$$

Since by construction $\neg\langle\Uparrow \varphi_1\rangle\top \rightarrow \langle\Uparrow \varphi_1\rangle(\langle\Uparrow \varphi_2\rangle\top \rightarrow (\dots\langle\Uparrow \varphi_n\rangle(\langle\Uparrow \varphi_{n+1}\rangle\top \rightarrow \langle\Uparrow \varphi_{n+1}\rangle\psi)\dots)) \in \lambda(v)$ it follows that $\langle\Uparrow \varphi_1\rangle\top \in \lambda(v)$ (12) and $\neg\langle\Uparrow \varphi_1\rangle(\langle\Uparrow \varphi_2\rangle\top \rightarrow (\dots\langle\Uparrow \varphi_n\rangle(\langle\Uparrow \varphi_{n+1}\rangle\top \rightarrow \langle\Uparrow \varphi_{n+1}\rangle\psi)\dots))$ (13). But it follows from (12), (13) and $(\langle\Uparrow\rangle\neg)$, that $\langle\Uparrow \varphi_1\rangle\neg(\langle\Uparrow \varphi_2\rangle\top \rightarrow (\dots\langle\Uparrow \varphi_n\rangle(\langle\Uparrow \varphi_{n+1}\rangle\top \rightarrow \langle\Uparrow \varphi_{n+1}\rangle\psi)\dots)) \in \lambda(v)$ (14). But then by construction $\neg(\langle\Uparrow \varphi_2\rangle\top \rightarrow (\dots\langle\Uparrow \varphi_n\rangle(\langle\Uparrow \varphi_{n+1}\rangle\top \rightarrow \langle\Uparrow \varphi_{n+1}\rangle\psi)\dots)) \in \lambda(v\varphi_1)$ (15). Repeating this argument we find that $\neg\psi \in \lambda(v\varphi_1 \dots \varphi_{n+1})$ (16). By Lemma 3.24 $\lambda(v\varphi_1 \dots \varphi_{n+1})$ is consistent, (16) therefore implies that $\psi \notin \lambda(v\varphi_1 \dots \varphi_{n+1})$ (17). But $(A_{\equiv})$, (11) and (17) is all we need to prove (7). We can now apply the main IH on formulas to get $\mathcal{H}^{\equiv}, v\varphi_1 \dots \varphi_{n+1} \not\Vdash \psi$ (18). By (18) and the truth conditions of $A_{\equiv}$ it follows that $\mathcal{H}^{\Sigma}, h \Vdash \mathtt{A}_{\equiv}\psi$. This concludes the proof for this direction of the $A_{\equiv}$-subcase.

The case for K_i is similar.

[$[\leq]$-modality.] *From left to right.* Assume that $[\leq_i]\psi \in \lambda(h)$ (0). There are two cases. Either $h \in H^0$ (1) or $h \in (H^{\Sigma} - H^0)$ (2).

Let us consider the first case. Assume that $h, h' \in H^0$ and that $h \leq^{\Sigma} h'$ (3). It follows from (3) by construction that $\{\varphi \mid [\leq]\varphi \in \lambda(h)\} \subseteq \lambda(h')$ (4). From (4) and (0) we know in particular that $\psi \in \lambda(h')$ (5). By (5) and the IH of the main induction on formulas it follows that $\mathcal{H}^{\Sigma}, h' \Vdash \psi$ (6). Since h' was arbitrary, it follows therefore from (6) and the truth definition of $[\leq]$ that $\mathcal{H}^{\Sigma}, h \Vdash [\leq]\psi$ (7).

Let us now consider the second case: $h \in (H^{\Sigma} - H^0)$ (2). For simplicity we assume that h is of the form $w\varphi$ (8). The proof can be generalized along the lines of the $\mathtt{A}_{\equiv}$-case. From (8) and (0) it follows by construction that $\langle\Uparrow \varphi\rangle[\leq]\psi \in \lambda(w\varphi)$ (9). Since by Lemma 3.24 $\lambda(w)$ is a $\langle\Uparrow\rangle$**DOX**-MCS, it follows from (9) and Axiom $(\langle\Uparrow\rangle[\leq_i])$ that $\langle\Uparrow \varphi_{n+1}\rangle\top \in \lambda(w)$ (10) and $[(\varphi \rightarrow \mathtt{A}_{\equiv}(\neg\varphi \rightarrow \neg\langle\Uparrow \varphi\rangle\neg\psi)) \wedge (\varphi \rightarrow [\leq_i]\neg\langle\Uparrow \varphi\rangle\neg\psi) \wedge [\leq_i](\neg\varphi \rightarrow \neg\langle\Uparrow \varphi\rangle\neg\psi)] \in \lambda(w)$ (11).

Now assume that $h \leq^{\Sigma} h'$ (15). It follows from (15) by construction that h' is of the form $v\varphi$ (16). By (15) we know that we are in one of the following cases:

1. $\varphi \in \lambda(w)$ while $\varphi \notin \lambda(v)$
2. $\varphi \in \lambda(w)$, $\varphi \in \lambda(v)$, and $w \leq_i v$
3. $\varphi \notin \lambda(w)$, $\varphi \notin \lambda(v)$, and $w \leq_i v$.

Case 1: $\varphi \in \lambda(w)$ (17.1) while $\varphi \notin \lambda(v)$ (17.2). By (11) we have: $(\varphi \rightarrow \mathtt{A}_{\equiv}(\neg\varphi \rightarrow \neg\langle\Uparrow \varphi\rangle\neg\psi)) \in w$ (17.3). It follows from (17.1), (17.3) and Lemma 3.24 that $(\mathtt{A}_{\equiv}(\neg\varphi \rightarrow \neg\langle\Uparrow \varphi\rangle\neg\psi)) \in w$ (17.4). From (17.4) we have by construction $(\neg\varphi \rightarrow \neg\langle\Uparrow \varphi\rangle\neg\psi) \in v$ (17.5). But it is easy to check that $\neg\varphi \in v$ (17.6). Thus by (17.5), (17.6) and Lemma 3.24 we have $\neg\langle\Uparrow \varphi\rangle\neg\psi \in v$ (17.7). Again it is easy to check that $\langle\Uparrow \varphi\rangle\top \in v$ (17.8). But (17.7), (17.8), $(\langle\Uparrow\rangle\neg)$ and Lemma 3.24 gives us $\langle\Uparrow \varphi\rangle\psi \in v$ (17.9). By construction (17.9) gives us $\psi \in v\varphi$ (17.10). (17.10) gives us by IH $h' \Vdash \psi$ (17.11). But since h' was arbitrary, we have: $\mathcal{H}^{\Sigma}, h \Vdash [\leq]\psi$ (17.12).

Case 2: $\varphi \in \lambda(w)$ (18.1), $\varphi \in \lambda(v)$ (18.2) and $w \leq v$ (18.3). By (11) we have: $(\varphi \rightarrow [\leq_i]\neg\langle\Uparrow \varphi\rangle\neg\psi) \in w$ (17.3). An easy argument gives us $[\leq_i]\neg\langle\Uparrow \varphi\rangle\neg\psi \in w$ (17.4). Thus by construction $\neg\langle\Uparrow \varphi\rangle\neg\psi \in v$ (17.5). By construction we have also $\langle\Uparrow \varphi\rangle\top \in v$ which together with (17.5) and $(\langle\Uparrow\rangle\neg)$ gives us $\langle\Uparrow \varphi\rangle\psi \in v$ (17.6). (17.6) implies by construction that $\psi \in v$. The usual argument concludes the proof.

Case 3: $\varphi \notin \lambda(w)$ (19.1), $\varphi \notin \lambda(v)$ (19.2) and $w \leq v$ (19.3). By (11) we have: $[\leq_i](\neg\varphi \rightarrow \neg\langle\Uparrow \varphi\rangle\neg\psi) \in w$ (19.4). An easy argument gives us $[\leq_i]\neg\langle\Uparrow \varphi\rangle\neg\psi \in w$ (17.4). Thus by construction $\neg\langle\Uparrow \varphi\rangle\neg\psi \in v$ (17.5). By construction we have also $\langle\Uparrow \varphi\rangle\top \in v$ which together with (17.5) and $(\langle\Uparrow\rangle\neg)$ gives us $\langle\Uparrow \varphi\rangle\psi \in v$ (17.6). (17.6) implies by construction that $\psi \in v$. The usual argument concludes the proof for this case and this direction.

[$[\leq_i]$-modality.] *From right to left.* Assume that $[\leq_i]\psi \notin \lambda(h)$ (0). There are two cases. Either $h \in H^0$ (1) or $h \in (H^{\Sigma} - H^0)$ (2).

The first case is along the lines of the proof in the previous section.

Let us now consider the other case. For the sake of simplicity we assume that $h = w\varphi$. The proof can be generalized along the lines of the $A_{\equiv}$-case. Now assume that $[\leq_i]\psi \notin \lambda(h)$ (0). It follows from (0) by maximality of $\lambda(h)$ and construction that $\langle\Uparrow \varphi\rangle\neg[\leq_i]\psi \in \lambda(w)$ (1). It follows from (1) by $(\langle\Uparrow\rangle\neg)$ that $\langle\Uparrow \varphi\rangle\top \in \lambda(w)$ (2) and $\neg\langle\Uparrow \varphi\rangle[\leq_i]\psi \in \lambda(w)$ (3).

Given that (2) and (3), an easy argument give us by $(\langle\Uparrow\rangle[\leq_i])$ that $\neg[(\varphi \rightarrow \mathtt{A}_{\equiv}(\neg\varphi \rightarrow \neg\langle\Uparrow \varphi\rangle\neg\psi)) \wedge (\varphi \rightarrow [\leq_i]\neg\langle\Uparrow \varphi\rangle\neg\psi) \wedge [\leq_i](\neg\varphi \rightarrow \neg\langle\Uparrow \varphi\rangle\neg\psi)] \notin w$ and thus that $\neg(\varphi \rightarrow \mathtt{A}_{\equiv}(\neg\varphi \rightarrow \neg\langle\Uparrow \varphi\rangle\neg\psi)) \vee \neg(\varphi \rightarrow [\leq_i]\neg\langle\Uparrow \varphi\rangle\neg\psi) \vee \neg[\leq_i](\neg\varphi \rightarrow \neg\langle\Uparrow \varphi\rangle\neg\psi)]$ (4). The preceding disjunction naturally displays three cases.

Case 1: $\neg(\varphi \rightarrow \mathtt{A}_{\equiv}(\neg\varphi \rightarrow \neg\langle\Uparrow \varphi\rangle\neg\psi)) \in w$ (4.1). Since by Lemma 3.24 $\varphi \in w$ (4.2) and $\neg\mathtt{A}_{\equiv}(\neg\varphi \rightarrow \neg\langle\Uparrow \varphi\rangle\neg\psi) \in w$ (4.3). We will now prove that there is a v such that $w \equiv v$, $\varphi \notin v$ and $\psi \notin \lambda(v\varphi)$. First take the following set $v_0 = \{\chi \mid A \equiv \chi \in \lambda(w)\} \cup \{\neg\varphi \wedge \langle\Uparrow \varphi\rangle\neg\psi\}$ (4.5). Assume for a contradiction that v_0 is inconsistent (4.6). It follows from (4.6) that there is a finite set of formulas $\{\chi_1, \ldots \chi_m\} \subseteq \{\chi \mid A \equiv \chi \in \lambda(w)\}$ such that $\vdash (\bigwedge_{i=1}^{m} \chi_i) \rightarrow (\neg\varphi \rightarrow \neg\langle\Uparrow \varphi\rangle\neg\psi)$ (4.7). But from (4.7) and standard modal reasoning we find that $\vdash (\bigwedge_{i=1}^{m} \mathtt{A}_{\equiv}\chi_i) \rightarrow \mathtt{A}_{\equiv}(\neg\varphi \rightarrow \neg\langle\Uparrow \varphi\rangle\neg\psi)$ (4.8). But from (4.8), (4.2) and Lemma 3.24 we have $\mathtt{A}_{\equiv}(\neg\varphi \rightarrow \neg\langle\Uparrow \varphi\rangle\neg\psi)$ (4.9). But (4.9) contradicts (4.3), thus by reduction v_0 is consistent. By the Lindenbaum Lemma it can be extended to a maximal consistent set v^+. But by construction $v^+ \in H^{\Sigma}$ (4.10). Since $\{\chi \mid A_{\equiv}\chi \in \lambda(w)\} \subseteq v_0 \subseteq v^+$ it follows by construction that $w \equiv v^+$ (4.11). By construction of v_0 we have also $\neg\varphi \in v_0 \subseteq v^+$ (4.12). But it follows from (4.2), (4.12) and (4.11) by construction that $w\varphi \leq v\varphi$ (4.13). But since by (4.5) and a simple argument $\langle\Uparrow \varphi\rangle\neg\psi \in v$ too, we have $\psi \notin v\varphi$ (4.14). But (4.13) and (4.14) gives us $w\varphi \not\Vdash [\leq_i]\psi$. This concludes our proof for this subcase.

Case 2: $\neg(\varphi \rightarrow [\leq_i]\neg\langle\Uparrow \varphi\rangle\neg\psi) \in w$ (5.1). It follows from (5.1) by Lemma 3.24 that $\varphi \in w$ (5.2) and that $\neg[\leq]\neg\langle\Uparrow \varphi\rangle\neg\psi \in w$ (5.3). By $(\langle\Uparrow\rangle\neg)$ it follows that $\neg[\leq](\langle\Uparrow \varphi\rangle\top \rightarrow \langle\Uparrow \varphi\rangle\psi) \in w$. The right to left argument of the $\mathtt{A}_{\equiv}$-case applies replacing $\equiv$ by $\leq$, i.e. we construct v^+ such that by construction $w \leq v$ (5.4) and $\langle\Uparrow \varphi\rangle\neg\psi \in v^+$ (5.5). We find by (5.5) and construction that $\psi \notin v\varphi$ (5.6). But since by construction (5.4) and (5.2) give us $w\varphi \leq v\varphi$ (5.7). The rest of the argument is as usual.

Case 3: $\neg[\leq_i](\neg\varphi \rightarrow \neg\langle\Uparrow \varphi\rangle\neg\psi) \in w$ (6.1). An easy argument show that by (6.1) we have $\neg[\leq_i](\neg\varphi \rightarrow (\langle\Uparrow \varphi\rangle\top \rightarrow \langle\Uparrow \varphi\rangle\psi)) \in w$ (6.2). We construct v_0 such that $\neg\varphi \in v_0$ (6.3), $\langle\Uparrow \varphi\rangle\top \in v_0$ (6.4) but $\langle\Uparrow \varphi\rangle\psi \notin v_0$ (6.5) that can be extended to a MCS v^+ (6.6) such that $w \leq v^+$ (6.7). By (6.6), (6.7) and (6.3) we have by construction $w\varphi \leq v\varphi$ (6.8). But from (6.3) and (6.8) an easy argument gives the desired result by the main IH. This concludes the proof for this subcase, this direction and the case.

The case of $\Box^{\geq}$ is similar. QED

We have shown that for every consistent set of formulas, there was an initial state in our canonical forest such that all formulas in the set were satisfied there. We finally need to show that our canonical forest can be generated by an initial plausibility model and a dynamic belief revision protocol.

Lemma 3.26. *Any subforest of the canonical forest generated by an initial state in the initial canonical epistemic plausibility model is isomorphic to a forest generated by an initial plausibility model and a dynamic belief revision protocol.*

Proof. We start by proving that the initial part of the canonical forest is isomorphic to an epistemic plausibility model $\mathcal{M}$ (1). This proof follows by a classical

Sahlqvist correspondence argument [39, ch. 4] and preservation of the relevant properties under taking generated subframes. The guarantee that $\equiv$ is the universal relation follows from taking a point-generated submodel (see [39, 7.1] for details). It remains to prove that there is a state-dependent dynamic belief revision protocol $\mathtt{p} : W \to Ptcl(\mathcal{L})$ such that the forest generated by the initial part of the canonical forest according to the protocol $\mathtt{p}$ is isomorphic to the canonical forest. Pick $\mathtt{p}$ such that $\mathtt{p}(w) = \{\sigma \mid w\sigma \in H^{\Sigma}\}$. We claim that a history $w\sigma \in H^{\Sigma}$ iff $h\sigma \in H(\mathcal{M}, p) = Dom(\mathcal{H}(\mathcal{M}, p))$. (The rest of the Lemma — i.e. the clauses for the plausibility relation and for the valuation function — follows by construction as the reader can check by inspecting the clauses in Definition 3.20 and Definition 3.23.) The proof is by induction on the length of $w\sigma$. The base case is immediate from (1). Now assume that the equivalence holds for histories of length n. Assume that $w\sigma\varphi$ is of length $n+1$ (2) and $w\sigma\varphi \in H^{\Sigma}$ (3). It follows by construction of the canonical model that $w\sigma \in H^{\Sigma}$ (4). But then by IH and (4) we have $w\sigma \in H(\mathcal{M}, p)$ (5). It follows by construction of $\mathtt{p}$ and (3) that $\sigma\varphi \in \mathtt{p}$ (6). But then by Definition 3.20, (5) and (6) we have $w\sigma\varphi \in H(\mathcal{M}, p)$ (7). The other direction is similar. QED

Theorem 3.27 (Completeness)**.** $\langle\Uparrow\rangle$**DOX** *is sound and strongly complete with respect to the class of DoTL models generated by lexicographically upgrading some doxastic model according to some state-dependent dynamic belief revision protocol.*

Proof. The proof follows from the Truth Lemma and the preceding Lemma by a standard argument [39]. QED

3.4.3 More languages

How can we extend the preceding completeness proof to axiomatize dynamic logics based on other static languages? As we can see from the previous proof this takes two separate task. First we should be able to construct a satisfactory initial canonical model and then extend it to a canonical forest. We have discussed strategies for builiding canonical models for a range of doxastic static languages at the end of Section 3.1. Let us now give dynamic-temporal axioms for other interesting modalities. First of all for safe belief $\Box_i\varphi$:

$$\begin{aligned}\langle\Uparrow\varphi\rangle\Box_i\psi \;\leftrightarrow\; \langle\Uparrow\varphi\rangle\top \wedge \\ (\neg\varphi \to \langle i\rangle(\varphi \to (\langle\Uparrow\varphi\rangle\top \to \langle\Uparrow\varphi\rangle\psi)) \\ \wedge\ (\neg\varphi \to \Box_i(\langle\Uparrow\varphi\rangle\top \wedge \langle\Uparrow\varphi\rangle\psi)) \\ \wedge\ (\varphi \to \Box_i(\varphi \to (\langle\Uparrow\varphi\rangle\top \to \langle\Uparrow\varphi\rangle\psi))\end{aligned} \tag{3.6}$$

and now for its dual $\langle \sim_i \cap \geq_i \rangle$:

$$\begin{aligned} \langle \Uparrow \varphi \rangle \langle \sim_i \cap \geq_i \rangle \psi \;\leftrightarrow\; \langle \Uparrow \varphi \rangle \top \wedge [& \\ & \neg\varphi \wedge \langle i \rangle (\varphi \wedge \langle \Uparrow \varphi \rangle \psi) \\ & \vee (\neg\varphi \wedge \langle \sim_i \cap \geq_i \rangle (\neg\varphi \wedge \langle \Uparrow \varphi \rangle \psi)) \\ & \vee (\varphi \wedge \langle \sim_i \cap \geq_i \rangle (\varphi \wedge \langle \Uparrow \varphi \rangle \psi))] \end{aligned} \tag{3.7}$$

Recalling the translation given in the proof of Fact 3.5, the dynamic axiom for conditional belief $B_i^{\psi}\varphi$ operators follows from the two previous ones. We conclude by giving an explicit axiom for the special case of (non-conditional) belief.

$$\begin{aligned} \langle \Uparrow \varphi \rangle B_i \psi \;\leftrightarrow\; \langle \Uparrow \varphi \rangle \top \wedge & \\ & [(\langle i \rangle (\varphi \wedge \langle \Uparrow \varphi \rangle \top) \wedge K_i (\langle \sim_i \cap \geq_i \rangle ((\varphi \wedge \langle \Uparrow \varphi \rangle \top) \wedge \\ & \Box_i ((\varphi \wedge \langle \Uparrow \varphi \rangle \top) \rightarrow \langle \Uparrow \varphi \rangle \psi)))) \\ & \vee (K_i (\varphi \rightarrow \neg \langle \Uparrow \varphi \rangle \top) \wedge K_i (\langle \sim_i \cap \geq_i \rangle ((\langle \Uparrow \varphi \rangle \top) \wedge \\ & \Box_i (\langle \Uparrow \varphi \rangle \top \rightarrow \langle \Uparrow \varphi \rangle \psi)))))] \end{aligned} \tag{3.8}$$

The preceding completeness proof together with these additional dynamic-temporal axioms gave, we hope, interesting insights into the effect of protocols on logics of belief revision.

3.5 Conclusion

The previous chapter determined the special constraints that capture agents operating with the 'local updates' of dynamic doxastic logic by giving structural representation theorems.

Major sources. The first point of departure of this chapter is found in the recent work on `TPAL` and its axiomatic completeness made by Hoshi and his collaboratiors in [106, 36]. The second point of departure is van Benthem [29] which develops dynamic logics of belief revision and particularly the logic of lexicographic upgrade, and also the sequence of papers by Baltag and Smets, notably [16], who pursue the idea of compositional analysis initiated in Baltag et al. [20] for the doxastic case. The third source is the work done by Girard and his collaborators in [86], and also by Board [40], on conditional doxastic logic, its expressive power and its axiomatic completeness.

Our main results. This chapter took the comparison between temporal logical and dynamic logical approaches to belief change from the structural to the syntactic level. We discussed different static doxastic languages and their relative expressive power and gave a compositional analysis for the dynamic logics of belief change built on the top of them. Finally we have developed a systematic 'protocol

logic' of axiomatic completeness for constrained revision processes, analogous to the purely epistemic theory of observation and conversation protocols initiated in van Benthem et al. [36] and Hoshi [106].

The next step. The next chapters show applications of these models and languages to particular types of reasoning at work in intelligent interaction that the idea of belief change underlies, such as interactive reasoning, inductive reasoning and strategic reasoning. Let us give more details on the rest of the program. In the next chapter (Chapter 4) we discuss agreement results (in the line of Aumann [13]) and their dynamic counterpart in the context of these qualitative structures, bringing a logical perspective on a foundational layer of interactive epistemology. Chapter 5 analyses inductive reasoning and develops connections with formal learning theory. Chapter 6 is an analysis of knowledge, belief and their dynamics in games in the line of [40, 31, 65, 21]. Finally Chapter 7 pursues the issues of logical design discussed here for beliefs and their dynamics by analyzing the expressive power required by game-theoretical and social-theoretical concepts involving the complementary notions of coalitional powers and of preferences. But let us start with interactive reasoning.

Chapter 4

Agreement Theorems in Dynamic-Epistemic Logic[1]

The previous chapters developed a logical framework to reason about agents' beliefs and their dynamics at the interface of two families of temporal and dynamic doxastic logics. From this new logical point of view we build our first connection toward a non-logical research field concerned with belief dynamics. This first connection is with interactive epistemology: the study of interactive or higher-order reasoning, one of the foundational layers of the epistemic program in game theory. An important and seminal question for interactive epistemology was whether agents whose differences in beliefs arise only from differences in the information they have received and not from prior beliefs can 'agree to disagree'. Taking 'agree' to mean common knowledge and 'disagree' to have different (posterior) beliefs about some event, Aumann [13] showed that this is impossible: common knowledge of disagreement implies differences in prior beliefs. This chapter contributes a logical perspective to the question.

4.1 Introduction

In this chapter we study Aumann's Agreement Theorem [13] and some of its subsequent extensions [82] and generalizations [57, 14] from the perspective of dynamic-epistemic logic [19, 67]. We show that common *belief* of posteriors is sufficient for agreement in 'epistemic-plausibility models', under common and well-founded priors, from which the usual form of agreement results follows, using common knowledge. Recent work [6, 7] has focused on epistemic foundations of solution concepts for games with possibly infinite strategy sets. In this line we do not restrict ourselves to the finite case, which also represents an improvement on known qualitative agreement theorems [14], and show that in countable structures

[1]This chapter is based on Dégremont and Roy [64].

such results hold if *and only if* the underlying 'plausibility ordering' is well-founded. We then look at these results from a syntactic point of view, showing that neither well-foundedness nor common priors are expressible in the language proposed in [17], even if it is extended with a common belief operator, but we also show a finitary syntactic derivation of the static agreement result in an extended modal language. We finally consider 'dynamic' agreement results. We show that 'agreements via dialogues' [57, 14] have a counterpart in epistemic-plausibility models, and that one also gets agreements via 'public announcements', a type of belief update that has so far not been considered in the agreement literature — see [49] and [121] for surveys. Comparison of the two types of dynamic agreements reveals that in some situations they are indeed different.

These technical results answer an 'internal' question in dynamic-epistemic logic, namely whether agreement results hold in this framework, but they also offer new insights into the contribution of agreement theorems to interactive epistemology. That common belief of posteriors is sufficient for agreement, under common and well-founded priors, strengthens one of the key lessons of agreement theorems, viz. that first-order information is closely dependent on higher-order information in situations of interaction [49]. Our inexpressibility results, on the other hand, support a qualm already voiced in the literature concerning the difficulty for agents to reason about static agreements [142]. The two dynamic results not only make a sharp distinction between two forms of belief changes; they also allow one to capture more adequately the idea that agreements are reached via *public* dialogues. Bringing agreement theorems to dynamic-epistemic logic is thus important both technically and conceptually, and it helps to bridge the existing literature on agreements with the logical approaches to knowledge, beliefs and the dynamics of information.

4.2 Definitions

In this section we introduce the models in which we study the various agreement results, and the logical language used in [17] to describe them. We assume the definitions given in Chapter 1 and especially in Section 1.3.3. We go again through some of them in order to draw connections with the probabilistic case on the way. For background on probabilistic interactive epistemology the reader can check Appendix A or for more details [127, ch.5]. We mentioned [49] and [121] for surveys of agreement theorems.

4.2.1 Epistemic plausibility models

We will be using (the at this point customary) epistemic plausibility models as a qualitative representation of the agents' beliefs as well as first- and higher-order information in a given interactive situation. Moreover we will assume that agents'

plausibility pre-orders are *total*.
The total plausibility pre-order $\leq_i$ induces i's *priors*, and can be viewed as a qualitative counterpart to a prior probability distribution on W. If $w \leq_i w'$ we say that i considers w at least as plausible as w'. Given a set $X \subseteq W$, we say that $w \in X$ is $\leq_i$-minimal in X if $w \leq_i w'$ for all $w' \in X$. The relation $\sim_i$ induces i's *information partition* of W. $\mathcal{K}_i[w]$ should be regarded as i's (private) information at w. We write $|\mathcal{M}| = W$ for the domain of $\mathcal{M}$ and use interchangeably I or N for the set of agents.

We remind the reader of the definition of two assumptions that are crucial in the following.

Definition 4.1 (Local well-foundedness)**.** *A plausibility pre-order satisfies:*

- **Local well-foundedness:** *If for all $w \in W$ and $i \in I$, for all $X \subseteq \mathcal{K}_i[w]$, if X is non-empty, then X has $\leq_i$-minimal elements.*
- **Well-foundedness:** *If for all $X \subseteq W$ and $i \in I$, if X is non-empty, then X has $\leq_i$-minimal elements.*

An epistemic plausibility model $\mathcal{M}$ satisfies (Local) Well-foundedness *if every plausibility pre-order has the corresponding property.*

Observe that β_i is well-defined if the plausibility pre-order is well-founded, while local well-foundedness is sufficient for $\mathcal{B}_i$ to be well-defined. To draw an analogy with the probabilistic case, this means that local well-foundedness ensures that the conditional beliefs of an agent i are well-defined for all 'events' that have a non-empty intersection with the agent's information partition. Well-foundedness, on the other hand, requires i's conditional beliefs to be well-defined for any non-empty subset of W. We now turn to the definition of the common prior (or same prior) assumption.

Definition 4.2 (Common Prior)**.** *There is* common prior beliefs *among group G in an epistemic plausibility model $\mathcal{M}$ when $\leq_i = \leq_j$ for all $i, j \in G$.*

The reflexive-transitive closure of the union of the epistemic accessibility relations $\sim_i$ for all agents i in a group G is the model-theoretic counterpart of the notion of 'common knowledge' in G [72, 67]. We define 'common belief' analogously.

Definition 4.3 (Common knowledge)**.** *For each $G \subseteq I$, let $\sim_G^*$ be the reflexive-transitive closure of $\bigcup_{i \in G} \sim_i$. Let $[w]_G^* = \{w' \in W \mid w \sim_G^* w'\}$.*

Definition 4.4 (Common belief)**.** *For each $G \subseteq I$, let $\rhd_G^*$ be the reflexive-transitive closure of $\bigcup_{i \in G} \rhd_i^{\mathcal{B}}$.*

4.2.2 Doxastic-epistemic logic

The logical language used in [17] to describe epistemic-plausibility models is a propositional modal language with three families of modal operators, which we extend here with 'common belief' operators. Note that it is an extension of the basic epistemic doxastic language $\mathcal{L}_{DOX}$ introduced in Section 1.3.3.

Definition 4.5 (Multi-Agent Epistemic Doxastic Language)**.** *The multi-agent epistemic doxastic language $\mathcal{L}_{EDL}$ is defined as follows:*

$$\varphi ::= p \mid \neg\varphi \mid \varphi \wedge \varphi \mid K_i\varphi \mid B_i^{\varphi}\varphi \mid C_G\varphi \mid CB_G\varphi,$$

where i ranges over I, p over a countable set of proposition letters PROP *and $\emptyset \neq G \subseteq I$.*

A formula $C_G\varphi$ is read as "it is common knowledge among group G that φ", $CB_G\varphi$ as "it is common belief among group G that φ." These formulas are interpreted in epistemic plausibility models as follows:

Definition 4.6 (Truth definition)**.**

$$\begin{array}{lll} \mathcal{M}, w \Vdash C_G\varphi & \text{iff} & \text{for all } v \text{ such that } w \sim_G^* v \text{ we have } \mathcal{M}, v \Vdash \varphi \\ \mathcal{M}, w \Vdash CB_G\varphi & \text{iff} & \text{for all } v \text{ such that } w \rhd_G^* v \text{ we have } \mathcal{M}, v \Vdash \varphi \end{array}$$

Before moving to our main agreement result, let us make a few simple observations about agreements in epistemic plausibility models.

4.2.3 Information, priors, posteriors and agreement

Accordingly this subsection looks at three simple situations that illustrate three simple facts about agreement in epistemic plausibility models. To begin with, agents can have the same priors (plausibility ordering) but different posteriors (beliefs), when the information they possess is different.

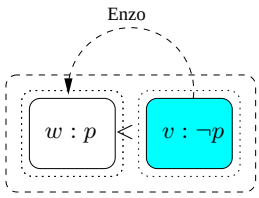

Figure 4.1: Same prior. Different posteriors.

In the model described in Figure 4.1, at v Enzo believes p, while Céline does not believe p (she even knows that $\neg p$), but they have the same prior. We will

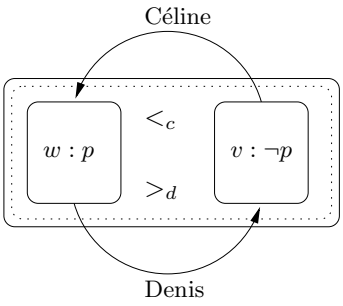

Figure 4.2: Common knowledge of disagreement. Different Priors.

now see that two agents can have different posteriors (disagreement) and that this fact can be common knowledge, when the agents have different priors.

In the model of Figure 4.2, at both w and v it is common knowledge between Céline and Denis that Denis believes p while Céline believes $\neg p$. We finally show that two agents with the same prior might have mutual knowledge of disagreement.

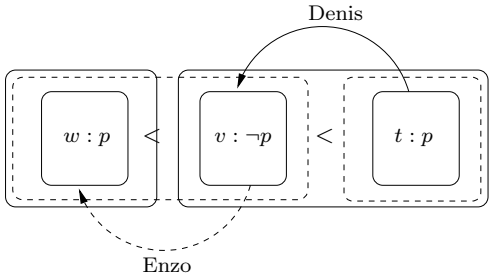

Figure 4.3: Mutual knowledge of disagreement. Same Prior.

Figure 4.3 represents a model in which, at t, Enzo knows that Denis and he disagree about p and so does Denis. How about higher levels of knowledge or belief? To start with, note that in this particular case it is no longer true for higher levels of knowledge or belief: indeed at t, Denis believes that Enzo believes that they both agree that p is the case. For a more general answer, keep reading.

4.3 Static agreement and well-foundedness

We first show that well-foundedness is sufficient for agreement on the posteriors under common priors and common *beliefs* of the posteriors. More precisely, we show that if an epistemic plausibility model is well-founded, then common belief that agent i believes that φ while j does not believe that φ implies that i and j have different priors, which is the contrapositive form of the agreement theorem.

Theorem 4.7 (Agreement theorem — Common Belief)**.** *If a* well-founded *epistemic plausibility model* $\mathcal{M}$ *satisfies* $\mathcal{M}, w \Vdash CB_{\{i,j\}}(B_i p \wedge \neg B_j p)$ *for some* $w \in W$*, then* i *and* j *have different priors in* $\mathcal{M}$*.*

Proof. We show that there is no pointed epistemic plausibility model $\mathcal{M}, w$ which satisfies *well-foundedness* and *common prior* such that $\mathcal{M}, w \Vdash CB_{\{i,j\}}(B_i p \wedge \neg B_j p)$. To do that we assume $\mathcal{M}, w \Vdash CB_{\{i,j\}}(B_i p \wedge \neg B_j p)$ and $\mathcal{M}$ satisfies *common prior* and show by induction that $\mathcal{M}$ must not be well-founded, by constructing an infinite descending chain $w_1 > w_2 > \ldots$, such that $w_1 \rhd^*_{\{1,2\}} w_n$ for every $n \in \omega$. Note that by common prior we have that $\leq_1 = \leq_2 = \leq$. Now assume that $\mathcal{M}, w \Vdash CB_{\{1,2\}}(B_1 p \wedge \neg B_2 p)$ (1) and suppose, towards a contradiction, that $\leq$ is well-founded.

Base case. We start by constructing a descending chain of length 2. By (1) we have in particular $\mathcal{M}, w \Vdash B_1(B_1 p \wedge \neg B_2 p)$. By assumption, it follows from the truth definition of B_1 (and $\leq_1 = \leq$-well-foundedness) that there is some state, call it w_0, such that $w_0 \in \min_{\leq} \mathcal{K}_1[w]$, i.e. $w \rhd^{\mathcal{B}}_i w_0$ (2) and $\mathcal{M}, w_0 \Vdash B_1 p \wedge \neg B_2 p$ (3). In particular $\mathcal{M}, w_0 \Vdash \neg B_2 p$ (4). By the same argument as before it follows that there must thus be a state, call it w_1, such that $w_1 \in \min_{\leq} \mathcal{K}_2[w_0]$, i.e. $w_0 \rhd^{\mathcal{B}}_2 w_1$ (5) and $\mathcal{M}, w_1 \Vdash \neg p$ (6).

But by (1), (2) and (5) it follows that $\mathcal{M}, w_1 \Vdash B_1 p$ (7), i.e. $\{v \in W | v \in \min_{\leq} \mathcal{K}_1[w_1]\} \subseteq V(p)$ (8). But then there is a state, call it w_2 such that $w_2 \in \min_{\leq} \mathcal{K}_1[w_1]$ (9) and $\mathcal{M}, w_2 \Vdash p$ (10). But then from (6), (8) and (9) it follows that $w_1 > w_2$.

Induction step. Assume that we have been able to construct a chain of length n, i.e. we have $w_1 > w_2 > \ldots > w_n$ such that $w_1 \rhd^*_{\{1,2\}} w_n$ (11) for every $n' \leq n$. Assume that there is no state v such that $w_n > v$ (12). Clearly w_n must be minimal within both $\mathcal{K}_1[w_n]$ (13) and $\mathcal{K}_2[w_n]$ (14). It is easy to see that by the truth condition of common belief we have by (2), (5), (11) and (1) that $\mathcal{M}, w_n \Vdash B_1 p$ (15). But then by (13) we have $w_n \Vdash p$ (16). Similarly we have $\mathcal{M}, w_n \Vdash \neg B_2 p$ (17). It then follows WLOG that there must be a state v_n such that $(v_n \leq w_n \;\&\; v_n \geq w_n)$ (18) such that $w_n \Vdash \neg p$ (19). It follows that $v_n \notin \mathcal{K}_1[w_n]$ (20). Moreover by common belief we have that $v_n \Vdash B_1 p$ (21). But it follows that $v_n \notin \min_{\leq} \mathcal{K}_1[v_n]$ (22). Since this set is non-empty it follows that there is some state $w_{n+1} \in \min_{\leq} \mathcal{K}_1[v_n]$ (23). But then we have by (22) and (23) that $v_n > w_{n+1}$ (24). But (24) and (18) implies that $w_n > w_{n+1}$ (25). This concludes the induction step and the proof. QED

This immediately implies the 'common knowledge' agreement result below, because $C_G \varphi \rightarrow CB_G \varphi$ is a valid implication in epistemic plausibility models.

Corollary 4.8 (Agreement theorem — Common Knowledge)**.** *If an epistemic plausibility model* $\mathcal{M}$ *satisfies* well-foundedness *and* $\mathcal{M}, w \Vdash C_{\{i,j\}}(B_i p \wedge \neg B_j p)$ *for one* $w \in W$*, then* i *and* j *have different priors in* $\mathcal{M}$*.*

Remark. As a side observation, note that it is possible to prove the corollary directly by application of Bacharach's [14] result on qualitative 'decision functions', modulo generalization to the countable case. But let us move back to the main line.

Well-foundeness is not only sufficient for common priors to exclude the possibility of disagreements when the posteriors are common beliefs, it is also necessary, as Proposition 4.9 shows. The model behind this result is drawn in figure 4.4.

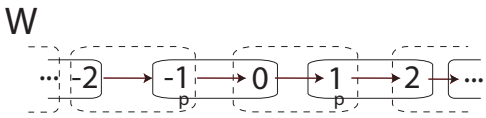

Figure 4.4: The epistemic plausibility model constructed in the proof of Proposition 4.9. The solid and dotted rectangles represent 1's and 2's information partitions on W, respectively. The arrows represent their common plausibility ordering.

Proposition 4.9. *There exists a pointed epistemic plausibility model* $\mathcal{M}, w$ *which satisfies* local well-foundedness *and* common prior *such that* $\mathcal{M}, w \Vdash C_{\{1,2\}}(B_1 p \wedge \neg B_2 p)$.

Proof. Let the model $\mathcal{M}$ be defined as follows, with $I = \{1, 2\}$.

- $\mathcal{M} = \langle \mathbb{Z}, (\leq_i)_{i \in I}, (\sim_i)_{i \in I}, V \rangle$ such that:
  - $\mathbb{Z}$ is the set of integers.
  - For both agents $i \in I$, $x \leq_i y$ iff $x \geq y$.
  - For all $x, y \in \mathbb{Z}$: $x \sim_1 y$ is the smallest equivalence relation such that $x \sim_1 y$ whenever $y = x + 1$ and x is odd; $x \sim_2 y$ is the smallest equivalence relation such that $x \sim_2 y$ whenever $y = x + 1$ and x is even.
  - $V(p) = \{x : x \text{ is odd } \}$ and $V(q) = \emptyset$ for all $q \neq p$ in PROP.

It is easily checked that at every $x \in \mathbb{Z}$ we have $\mathcal{M}, x \Vdash (\neg B_1 p \wedge B_2 p)$, and so that $\mathcal{M}, x \Vdash C_{\{1,2\}}(\neg B_1 p \wedge B_2 p)$, and moreover that $\mathcal{M}$ satisfies *local well-foundedness* and *common prior*. QED

To sum up, well-foundedness is thus sufficient for agreement results to hold, and furthermore cannot be weakened to *local well-foundedness*. This condition on the plausibility ordering is thus *the* safeguard against common knowledge of disagreement, once we drop the assumption that the state space is finite.

4.4 Expressive power and syntactic proofs

$\mathcal{L}_{EDL}$ is a natural choice of language for talking about epistemic-plausibility models, but we show here that it cannot express Theorem 4.7 nor Corollary 4.8, because it cannot express two of their key assumptions, common prior and well-foundedness.

Fact 4.10. *The class of epistemic plausibility frames that satisfies common prior is not definable in* $\mathcal{L}_{EDL}$.

Figure 4.5: The two epistemic plausibility model constructed in the proof of Fact 4.10. 1's and 2's information partitions on W are represented as in figure 4.4. The arrow in W represents their common plausibility ordering, while in W' the solid arrow and dotted arrows represent 1's and 2's orderings, respectively.

Proof. Take $W = \{x, y\}$ and $\sim_1 = \sim_2 = \{(x, x), (y, y)\}$. We consider two epistemic plausibility frames $\mathcal{F}$ and $\mathcal{F}'$. $\mathcal{F} = \langle W, \sim_1, \sim_2, \leq_1, \leq_2\rangle$ and $\mathcal{F}' = \langle W, \sim_1, \sim_2, \leq'_1, \leq'_2\rangle$ where $\leq_1 = \leq_2 = \{(x, x), (x, y), (y, y)\}$ while $\leq'_1 = \{(x, x), (x, y), (y, y)\}$ and $\leq'_2 = \{(x, x), (y, x), (y, y)\}$. Clearly $\mathcal{F}$ satisfies common prior while $\mathcal{F}'$ does not. Now assume for a contradiction that there is a formula $\psi \in \mathcal{L}_{EDL}$ that defines the class of epistemic plausibility frames with common prior. Then we have $\mathcal{F} \Vdash \psi$ (1) while $\mathcal{F}' \nVdash \psi$ (2). It follows from (2) that there is some valuation V and some state $s \in W = \{x, y\}$, such that $\mathcal{F}', V, s \nVdash \psi$ (3). But it follows from (1) that $\mathcal{F}, V, s \Vdash \psi$ (4).

We now prove by induction on the complexity of φ that for all $\varphi \in \mathcal{L}_{EDL}$ and for all $s \in W$ we have $\mathcal{F}, V, s \Vdash \varphi$ iff $\mathcal{F}', V, s \Vdash \varphi$ which together with (3) and (4) gives us a contradiction. The base case for propositional letters is immediate. The cases for common knowledge and knowledge follow from the fact that the pointed models $\mathcal{F}, V, s$ and $\mathcal{F}', V, s$ are isomorphic with respect to $\sim_i$. The case for common belief is trivial due to the fact that the information partitions are (isomorphic) singletons. Moreover the structures are fully isomorphic for agent 1. So it remains to consider the case of conditional belief for 2.

Take the state x (the proof is similar for y). Now assume that $\mathcal{F}, V, x \Vdash B_2^{\varphi}\chi$ (5). We need to show that $\mathcal{F}', V, x \Vdash B_2^{\varphi}\chi$. By IH we have $||\varphi||^{\mathcal{M}} = ||\varphi||^{\mathcal{M}'}$ (6), with $\mathcal{M} = \mathcal{F}, V$ and $\mathcal{M}' = \mathcal{F}', V$. Now (5) iff $\forall v$ (if $v \in \beta_2(\mathcal{K}_2[x] \cap ||\varphi||^{\mathcal{M}})$ then $\mathcal{M}, v \Vdash \chi$) (7). Observe that by (6) we know that $\beta_2(\mathcal{K}_2[x] \cap ||\varphi||^{\mathcal{M}}) = \beta'_2(\mathcal{K}_2[x] \cap ||\varphi||^{\mathcal{M}'})$ (8), since $\mathcal{K}_2[x] = \{x\}$ in $\mathcal{M}$ and $\mathcal{M}'$. Now if $(\mathcal{K}_2[x] \cap ||\varphi||^{\mathcal{M}}) = \emptyset$ we are done, and otherwise we use (7), (8), truth condition of $B_2^{\varphi}\chi$ and IH for χ. This concludes the induction step and the proof. QED

This result, which rests on the two small models drawn in figure 4.5, confirms the idea that to reason about (common) priors the agents must make "inter-[information]-state comparisons" [142], which they cannot do because their reasonings in $\mathcal{L}_{EDL}$ are local, i.e. they are bounded by the 'hard information' [29] they have. This limitation also makes well-foundeness inexpressible, and with it the two static agreement results.

Fact 4.11. *There is no formula φ of $\mathcal{L}_{EDL}$ which is true in a pointed epistemic plausibility model $\mathcal{M}, w$ iff Theorem 4.7 or Corollary 4.8 holds in $\mathcal{M}, w$.*

Proof. See Appendix D. QED

The previous facts teach us that the the syntactical counterparts of the model-theoretic agreement results thus reside in more expressive languages.

In what follows we present a finite syntactic derivation of Corollary 4.8 in the hybrid language $\mathcal{H}(@, \downarrow, \geq_j, \sim_j)$ with a common knowledge modality C_G, and give additional facts about this logic of agreement. (In fact we use a language with more primitives, but, as we will prove, these are entirely definable in the restricted language.) The language we use authorizes the following basic programs.

$$\alpha ::= 1 \mid 2 \mid 1 \cup 2 \mid (1 \cup 2)^* \mid \geq_j \mid >_j$$

where j ranges over $\{1, 2\}$. We additionally authorize intersection of the basic programs only (not of arbitrary programs).

$$\beta ::= \alpha \mid \alpha \cap \alpha$$

Finally we recursively define our language as follows:

$$\varphi ::= p \mid i \mid x \mid \neg\varphi \mid \varphi \wedge \varphi \mid \langle\beta\rangle\varphi \mid @_i\varphi \mid @_x\varphi \mid \downarrow x.\varphi$$

where i ranges over a countable set of nominals NOM, x over a countable set of state variables SVAR and p over a countable set of proposition letters PROP. All these sets are assumed to be disjoint. Let us call this language $\mathcal{H}(\downarrow, @)[1, 2, (1 \cup 2), \geq_j, >_j, C_G, Res(\cap)]$. We immediately stress that our usage of intersection, union and of the strict modality does not increase the expressive power beyond $\mathcal{H}(\downarrow, @)[1, 2, \geq_j, C_G]$, i.e. the fragment that does not allow intersection, union, or the strict modality. We give below reduction axioms that sustain this claim. But before that let us finish defining our language by giving its semantics.

The programs $\{1, 2, \geq_j, >_j\}$ are interpreted in the obvious way. For example R_1 stands for $\sim_1$. The language $\mathcal{H}(\downarrow, @)[1, 2, (1 \cup 2), \geq_j, >_j, C_G, Res(\cap)]$ is interpreted on epistemic plausibility models together with an assignment function $g : \text{SVAR} \to W$ that maps state variables to states. The valuation function

maps elements of NOM to singleton sets of states. The following clauses cover the interpretation of the binder, of state variables and of the @ operator.

$$\begin{array}{lll}
\mathcal{M}, g, w \Vdash x & \text{iff} & g(x) = w \\
\mathcal{M}, g, w \Vdash i & \text{iff} & w \in V(i) \\
\mathcal{M}, g, w \Vdash @_x\varphi & \text{iff} & \mathcal{M}, g, g(x) \Vdash \varphi \\
\mathcal{M}, g, w \Vdash @_i\varphi & \text{iff} & \mathcal{M}, g, v \Vdash \varphi \text{ where } V(i) = \{v\} \\
\mathcal{M}, g, w \Vdash \downarrow x.\varphi & \text{iff} & \mathcal{M}, g[g(x) := w], w \Vdash \varphi
\end{array}$$

For the basic modalities $\{1, 2, \geq_j, >_j\}$ we have the classical scheme:

$$\mathcal{M}, g, w \Vdash \langle\alpha\rangle\varphi \quad \text{iff} \quad \text{for some } v \text{ with } wR_\alpha v \text{ we have } \mathcal{M}, g, v \Vdash \varphi$$

For the fragment of PDL we are using the clauses are:

$$\begin{array}{lll}
\mathcal{M}, g, w \Vdash \langle 1 \cup 2\rangle\varphi & \text{iff} & \exists v \text{ with } (w \sim_1 v \text{ or } w \sim_2 v) \text{ such that } \mathcal{M}, g, v \Vdash \varphi \\
\mathcal{M}, g, w \Vdash \langle (1 \cup 2)^*\rangle\varphi & \text{iff} & \exists v \text{ with } w \sim^*_{\{1\cup 2\}} v \text{ such that } \mathcal{M}, g, v \Vdash \varphi
\end{array}$$

The second operator is nothing but a notational variation of common knowledge, in the sense that $C_{\{1,2\}}\varphi \leftrightarrow \neg\langle(1 \cup 2)^*\rangle\neg\varphi$ which is useful to shorten our formulas when intersection is involved. The first one is the diamond version of the 'everybody knows' modality. Finally we give the clause for intersection.

$$\mathcal{M}, g, w \Vdash \langle\alpha \cap \beta\rangle\varphi \quad \text{iff} \quad \exists v \text{ with } (wR_\alpha v \text{ and } wR_\beta v) \text{ such that } \mathcal{M}, g, v \Vdash \varphi$$

We now show that the hybrid language $\mathcal{H}(@, \downarrow, \geq_j, \sim_j)$ with a common knowledge modality C_G is actually as expressive as the previous language.

Proposition 4.12. *The following reduction axioms are sound on the class of epistemic plausibility models.*

1. $\langle>\rangle\varphi \leftrightarrow \downarrow x.\langle\geq\rangle(\varphi \wedge [\geq]\neg x)$ *where* x *does not occur in* φ.
2. $\langle\alpha \cap \beta\rangle\varphi \leftrightarrow \downarrow x.\langle\alpha\rangle(\downarrow y.(\varphi \wedge @_x\langle\beta\rangle y))$ *where* x, y *does not occur in* φ.
3. *For* $\alpha \in \{1, 2, (1 \cup 2), (1 \cup 2)^*, \geq_1, \geq_2\}$ *and* $\beta \in \{1, 2, (1 \cup 2), (1 \cup 2)^*\}$*:* $\langle\alpha \cap \beta\rangle\varphi \leftrightarrow \downarrow x.\langle\alpha\rangle(\varphi \wedge \langle\beta\rangle x))$ *where* x *does not occur in* φ.
4. *For* $\alpha \in \{1, 2, (1 \cup 2), (1 \cup 2)^*\}$*:* $\langle>_j \cap\alpha\rangle\varphi \leftrightarrow \downarrow x.\langle\geq\rangle(\varphi \wedge [\geq]\neg x \wedge \langle\alpha\rangle x))$ *where* x *does not occur in* φ.

Proof. The proofs are standard correspondence arguments. (2) is valid for arbitrary programs on relational structures. (3) and (4) use the symmetry of the epistemic relation. QED

Let us note that the latter reduction axioms, which draw on the symmetry of the epistemic relations, are more efficient in terms of hybrid operator alternation and the number of fresh variables we need. Therefore we will rather work with them in the syntactic proof.

Corollary 4.13. *On the class of epistemic plausibility models* $\mathcal{H}(\downarrow, @)[1, 2, \geq_j, C_G]$ *is at least as expressive as* $\mathcal{H}(\downarrow, @)[1, 2, (1 \cup 2) \geq_j, >_j, C_G, Res(\cap)]$.

In addition to the reduction axioms given and the axiomatization of the pure hybrid logic $\mathcal{H}(\downarrow, @)$ (see [55]) we will make us of additional axioms in this proof. Their soundness is proved below.

Proposition 4.14. *The following axioms are valid on the class of well-founded epistemic-plausibility models.*

5. $[>]([>]p \rightarrow p) \rightarrow [>]p$
6. $\downarrow x.((\langle\geq\rangle(\neg\langle\geq\rangle x \wedge p)) \rightarrow \quad ((\langle\geq\rangle((\neg\langle\geq\rangle x \wedge p) \wedge \neg\downarrow z.\langle\geq\rangle(\neg\langle\geq\rangle z \wedge p)))))$
7. *For* $\alpha \in \{1, 2, (1 \cup 2), (1 \cup 2)^*\}$*:* $\downarrow x.[\alpha]\langle\alpha\rangle x$
8. $\langle\alpha \cap \beta\rangle\varphi \rightarrow (\langle\alpha\rangle\varphi \wedge \langle\beta\rangle\varphi)$
9. $\langle\alpha^*\rangle\varphi \leftrightarrow (\varphi \vee \langle\alpha\rangle\langle\alpha^*\rangle\varphi)$

Proof. (5) is sound on the class of $<$-well-founded frames (Löb axiom, see [39]). For (6) note that by Ax.(1), (6) is equivalent to $\langle>\rangle p \rightarrow \langle>\rangle(p \wedge \neg\langle>\rangle p)$ which is equivalent on the level of frames to (5). (7) is sound on the class of frames for which R_α is symmetric. (8) is obvious. (9) is the standard fixed point axiom of PDL. For all these facts see [39]. QED

We have now a language with sufficient expressive power to prove the syntactic counterpart to our agreement results for the case of common knowledge. In other words to show that agreement results can be recovered syntactically as theorems of a sound Hilbert axiom system for a multi-agent doxastic epistemic logic. Formally we prove the following:

Theorem 4.15. $\neg C_{\{1,2\}}(B_1 p \wedge \neg B_2 p)$ *is a theorem of the logic of* $\mathcal{H}(\downarrow, @)[\geq_j]$ *extended by Löb's axiom (items 5 and 6 in Proposition 4.14), the multi-agent* **S5***-epistemic logic including* C_G*,* **S4** *for* $\geq_j$ *and the Axiom of Common Prior:* $\langle\geq_i\rangle\varphi \leftrightarrow \langle\geq_j\rangle\varphi$.

Proof. For convenience we additionally use axioms 7 and 8 from Proposition 4.14 as useful shortcuts. (Löb($>$)) is either Axiom 5 or 6 in Proposition 4.14. For the axiomatization of $\mathcal{H}(\downarrow, @)[\geq_j]$ see [55]. For the multi-agent S5-epistemic logic including C_G see Section 1.3.2 or [72].

In the following proof after the axiom of common prior has been applied, we drop the label, since the plausibility relation is thus the same for both agents. Our goal is to derive a contradiction from the assumption that disagreement is common knowledge.

(0) $[(1 \cup 2)^*]\langle \geq \cap 2\rangle(([> \cap 2]\bot) \wedge \neg p)$ by Hypothesis.
(1) $[(1 \cup 2)^*][\geq \cap 1](([> \cap 1]\bot) \to p)$ by Hypothesis.
(2) $[(1 \cup 2)^*][2][\geq \cap 1](([> \cap 1]\bot) \to p)$ From (0) by PDL.
(3) $[(1 \cup 2)^*][\geq \cap 2][\geq \cap 1](([> \cap 1]\bot) \to p)$ From (2) by $\cap$.
(4) $[(1 \cup 2)^*][\geq \cap 2][\geq \cap 1](\neg p \to \top)$ From (3) by PL.
(4′) $[(1 \cup 2)^*][\geq \cap 2](\neg p \to \langle > \cap 1\rangle\top)$ From (3) by Ref for $(\geq \cap 1)$.
(5) $[(1 \cup 2)^*]\langle \geq \cap 2\rangle(([> \cap 2]\bot) \wedge \neg p \wedge (\neg p \to \langle > \cap 1\rangle\top))$ From (1) and (4') by ML.
(6) $[(1 \cup 2)^*]\langle \geq \cap 2\rangle(([> \cap 2]\bot) \wedge \neg p \wedge \langle > \cap 1\rangle\top)$ From (5) by PL and ML.

Let us now start a new derivation at the end of which we put (6) to use.

(7) $\downarrow x.((\langle\geq\rangle(\neg\langle\geq\rangle x \wedge p)) \to ((\langle\geq\rangle((\neg\langle\geq\rangle x \wedge p) \wedge \neg\downarrow z.\langle\geq\rangle(\neg\langle\geq\rangle z \wedge p)))))$
Axiom. Löb for $>$

(8) $\downarrow x.((\langle\geq\rangle(\neg\langle\geq\rangle x \wedge \langle(1 \cup 2)^*\rangle x)) \to ((\langle\geq\rangle((\neg\langle\geq\rangle x \wedge \langle(1 \cup 2)^*\rangle x) \wedge \neg\downarrow z.\langle\geq\rangle(\neg\langle\geq\rangle z \wedge \langle(1 \cup 2)^*\rangle x)))))$
From (7) by Uni Sub of p by $\langle(1 \cup 2)^*\rangle x$.

The previous step is conceptually important. (7) says that for every subset P of the domain, there is some minimal element with respect to the plausibility ordering. And (8) fixes P to be the 'common knowledge partition' we are in. It is safe since all p's are in the scope of the same binder for x (and x is the only free state variable in $\langle(1 \cup 2)^*\rangle x$).

(9) $[(1 \cup 2)^*][\geq \cap 2]\downarrow x.\big(((\langle\geq\rangle(\neg\langle\geq\rangle x \wedge \langle(1 \cup 2)^*\rangle x)) \to ((\langle\geq\rangle((\neg\langle\geq\rangle x \wedge \langle(1 \cup 2)^*\rangle x) \wedge \neg\downarrow z.\langle\geq\rangle(\neg\langle\geq\rangle z \wedge \langle(1 \cup 2)^*\rangle x)))))\big)$
From (8) by Necessitation

We start a new branch here in order to derive (12) and from it (13) using (6). (9) will be used to derive (15).

(10) $\langle > \cap 1\rangle p \leftrightarrow \downarrow x.\langle\geq\rangle((\neg\langle\geq\rangle x \wedge \langle 1\rangle x) \wedge p)$ Reduction Axiom $> \cap 1$
(11) $\langle > \cap 1\rangle\top \leftrightarrow \downarrow x.\langle\geq\rangle(\neg\langle\geq\rangle x \wedge \langle 1\rangle x)$ From (10) by Uni Sub of p by $\top$ and PL.
(12) $[(1 \cup 2)^*][\geq \cap 2](\langle > \cap 1\rangle\top \leftrightarrow \downarrow x.\langle\geq\rangle(\neg\langle\geq\rangle x \wedge \langle 1\rangle x))$ From (11) by Necessitation
(13) $[(1 \cup 2)^*]\langle \geq \cap 2\rangle(((([> \cap 2]\bot) \wedge \neg p \wedge \langle > \cap 1\rangle\top) \wedge \downarrow x.\langle\geq\rangle(\neg\langle\geq\rangle x \wedge \langle 1\rangle x))$
From (6) and (12) by ML.

We need a last step before we can prove (15).

(14) $\langle 1\rangle x \to \langle(1 \cup 2)^*\rangle x$ PDL
(15) $[(1 \cup 2)^*]\langle \geq \cap 2\rangle$ $((([> \cap 2]\bot) \wedge \neg p \wedge \langle > \cap 1\rangle\top) \wedge \downarrow x.[\ (\langle\geq\rangle(\neg\langle\geq\rangle x \wedge \langle 1\rangle x)) \wedge (\langle\geq\rangle\ (\neg\langle\geq\rangle x \wedge \langle(1 \cup 2)^*\rangle x) \wedge \neg\downarrow z.\langle\geq\rangle(\neg\langle\geq\rangle z \wedge \langle(1 \cup 2)^*\rangle x)\ )\ ]\ \ )$
From (13), (14) and (9) by ML and PL.
(16) $[(1 \cup 2)^*]\langle \geq \cap 2\rangle$ $((([> \cap 2]\bot) \wedge \neg p \wedge \langle > \cap 1\rangle\top) \wedge \downarrow x.[\ (\langle\geq\rangle(\neg\langle\geq\rangle x \wedge \langle 1\rangle x)) \wedge (\langle\geq\rangle\ (\neg\langle\geq\rangle x \wedge \langle(1 \cup 2)^*\rangle x) \wedge \downarrow z.[\geq](\langle(1 \cup 2)^*\rangle x \to \langle\geq\rangle z)\ )\ ]\ \ )$ From (15) by ML and PL.
(17) $[(1 \cup 2)^*]\langle \geq \cap 2\rangle$ $(\downarrow x.[(\langle\geq\rangle\ (\langle(1 \cup 2)^*\rangle x) \wedge \downarrow z.[\geq](\langle(1 \cup 2)^*\rangle x \to \langle\geq\rangle z)\ )\ ]\ \ )$ From (16) by ML and PL.

We start a new branch in order to derive (20).

(18)	$\downarrow x.(\langle\geq\rangle(\langle(1\cup 2)^*\rangle x \wedge p) \rightarrow \langle\geq\cap(1\cup 2)^*\rangle p)$	Axiom. By symmetry of $(1\cup 2)^*$ and $\cap$
(19)	$\downarrow x.(\langle\geq\rangle(\langle(1\cup 2)^*\rangle x \wedge$ $(\downarrow z.[\geq](\langle(1\cup 2)^*\rangle x \rightarrow \langle\geq\rangle z))) \rightarrow$ $\langle\geq\cap(1\cup 2)^*\rangle\downarrow z.[\geq](\langle(1\cup 2)^*\rangle x \rightarrow \langle\geq\rangle z))$	From (18) by Uni Sub of p by $\downarrow z.[\geq](\langle(1\cup 2)^*\rangle x \rightarrow \langle\geq\rangle z)$.
(20)	$[(1\cup 2)^*]\langle\geq\cap 2\rangle$ ($\downarrow x.\langle\geq\cap(1\cup 2)^*\rangle(\downarrow z.[\geq](\langle(1\cup 2)^*\rangle x \rightarrow \langle\geq\rangle z)$)	From (17) and (19) by ML and PL.

We leave (20) for now, start a new branch and use it to derive 26.

(21)	$[(1\cup 2)^*][(1\cup 2)^*][\geq\cap 1](([>\cap 1]\bot) \rightarrow p)$	From (1) by PDL.
(22)	$[(1\cup 2)^*][2][(1\cup 2)^*][2][\geq\cap 1](([>\cap 1]\bot) \rightarrow p)$	From (21) by PDL.
(23)	$[(1\cup 2)^*][2\cap\geq][(1\cup 2)^*\cap\geq][2\cap\geq][\geq\cap 1](([>\cap 1]\bot) \rightarrow p)$	From (20) by $\cap$
(24)	$[(1\cup 2)^*][2\cap\geq][(1\cup 2)^*\cap\geq][2\cap\geq][\geq\cap 1](\neg p \rightarrow (\langle>\cap 1\rangle\top))$	From (23) by PL and ML.

By a similar derivation we can derive (25) from (0).

(25)	$[(1\cup 2)^*][2\cap\geq][(1\cup 2)^*\cap\geq]\langle\geq\cap 2\rangle((([>\cap 2]\bot) \wedge \neg p)$	From (0) by a similar derivation.
(26)	$[(1\cup 2)^*]\langle\geq\cap 2\rangle$ ($\downarrow x.\langle\geq\cap(1\cup 2)^*\rangle[$ $(\downarrow z.[\geq](\langle(1\cup 2)^*\rangle x \rightarrow \langle\geq\rangle z)$) $\wedge$ $\langle\geq\cap 2\rangle((([>\cap 2]\bot) \wedge \neg p)]$	From (20) and (25) by ML and PL.

With (26) proved, we are now ready to head towards a contradiction. We start a new branch with a sequence of axioms.

(27)	$\downarrow x.[(1\cup 2)^*]\langle(1\cup 2)^*\rangle x$	Axiom for symmetry.
(28)	$\downarrow x.[(1\cup 2)^*][2]\langle(1\cup 2)^*\rangle x$	PDL.
(29)	$\downarrow x.[(1\cup 2)^*][2\cap\geq]\langle(1\cup 2)^*\rangle x$	By $\cap$
(30)	$[(1\cup 2)^*]\langle\geq\cap 2\rangle$ ($\downarrow x.\langle\geq\cap(1\cup 2)^*\rangle[$ $(\downarrow z.[\geq\cap 2](\langle(1\cup 2)^*\rangle x \rightarrow \langle\geq\rangle z)$)	From (26) by ML and $\cap$
(31)	$[(1\cup 2)^*]\langle\geq\cap 2\rangle$ ($\downarrow x.\langle\geq\cap(1\cup 2)^*\rangle[$ $(\downarrow z.\langle\geq\cap 2\rangle([$ $((([>\cap 2]\bot) \wedge \neg p) \wedge$ $(\langle(1\cup 2)^*\rangle x \wedge \langle\geq\rangle z)\ ]$	From (26), (29) and (30) by ML and PL.
(32)	$[(1\cup 2)^*]\langle\geq\cap 2\rangle$ ($\downarrow x.\langle\geq\cap(1\cup 2)^*\rangle[$ $(\downarrow z.\langle\geq\cap 2\rangle([$ $(\langle(1\cup 2)^*\rangle x \wedge \langle\geq\rangle z \wedge \langle>\cap 1\rangle\top) \wedge$ $([>\cap 2]\bot) \wedge \neg p)]$	From (26), (29) and (30) by ML and PL.

The final part of the derivation will use (26) and (32).

(33) $[(1 \cup 2)^*]\langle \geq \cap 2\rangle$ (
$\downarrow x.\langle \geq \cap (1 \cup 2)^*\rangle[$
$(\downarrow z.\langle \geq \cap 2\rangle[$
$(\langle > \cap 1\rangle \top))$ From (32) by ML and PL.

(34) $[(1 \cup 2)^*]\langle \geq \cap 2\rangle$ (
$\downarrow x.\langle \geq \cap (1 \cup 2)^*\rangle[$
$(\downarrow z.\langle \geq \cap 2\rangle[$
$(\langle \geq \cap 1\rangle(\neg\langle \geq\rangle z)))$ From (33) by ML and PL.

(35) $[(1 \cup 2)^*]\langle \geq \cap 2\rangle$ (
$\downarrow x.\langle \geq \cap (1 \cup 2)^*\rangle[$
$(\downarrow z.[\geq](\neg\langle \geq\rangle z \rightarrow \neg\langle (1 \cup 2)^*\rangle x) \quad)]$ From (26) by ML and PL.

(36) $[(1 \cup 2)^*]\langle \geq \cap 2\rangle$ (
$\downarrow x.\langle \geq \cap (1 \cup 2)^*\rangle[$
$(\downarrow z.\langle \geq \cap 2\rangle[$
$(\langle \geq \cap 1\rangle(\neg\langle (1 \cup 2)^*\rangle x)))$ From (34) and (35) by ML and PL.

(37) $[(1 \cup 2)^*]\langle 2\rangle$ (
$\downarrow x.\langle (1 \cup 2)^*\rangle[$
$(\downarrow z.\langle 2\rangle[$
$(\langle 1\rangle(\neg\langle (1 \cup 2)^*\rangle x)))$ From (36) by $\cap$.

(38) $[(1 \cup 2)^*]\langle 2\rangle$ (
$\downarrow x.\langle (1 \cup 2)^*\rangle(\neg\langle (1 \cup 2)^*\rangle x)))$ From (37) by PDL. (A contradiction)

QED

On the positive side we have seen that the hybrid language $\mathcal{H}(@, \downarrow, C_G, \geq_j, \sim_j)$ is able to axiomatize (converse) well-foundedness of the plausibility relation. On the negative side, the satisfiability problem for this language on the class of conversely well-founded frames is Σ^1_1-hard [55], ruling out any finite axiomatization of its validities. The derivation we show, however, is finite and uses only sound axioms. At the time of writing we still do not know whether the agreement results of Section 4.3 could be derived in a language whose validities over well-founded epistemic plausibility models are recursively enumerable. The fact that the syntactic derivation reported here pertains to such an expressive language nevertheless indicates that reasoning explicitly about agreement results requires onerous expressive resources.

4.5 Agreement via dialogues

In this section we turn to 'agreement-via-dialogues' [82, 14], which analyze how agents can reach agreement in the process of exchanging information about their beliefs by updating the latter accordingly.

4.5.1 Agreement via conditioning

We first consider agreements by repeated belief conditioning. It is known that if agents repeatedly exchange information about each others' posterior beliefs

about a certain event, and update these posteriors accordingly, the posteriors will eventually converge [82, 14]. We show here that this result also holds for the 'qualitative' form of belief conditionalization in epistemic plausibility models.

We call a *conditioning dialogue about* φ [82], at a state w of an epistemic plausibility model $\mathcal{M}$, a sequence of belief conditioning, for each agent, on all other agents' beliefs about φ. This sequence can be intuitively described as follows. It starts with the agents' simple belief about φ, i.e. for all i: $B_i\varphi$ if $\mathcal{M}, w \Vdash B_i\varphi$ and $\neg B_i\varphi$ otherwise. Agent i's belief about φ at the next stage is defined by taking his belief about φ, conditional upon learning the others' beliefs about φ at that stage. Syntactically, this gives, $\mathbb{B}_{1,i} = B_i\varphi$ if $\mathcal{M}, w \Vdash B_i\varphi$ and $\mathbb{B}_{1,i} = \neg B_i\varphi$ otherwise and, for two agents i, j, $\mathbb{B}_{n+1,i} = B_i^{\mathbb{B}_{n,j}\varphi}\varphi$ if $\mathcal{M}, w \Vdash B_i^{\mathbb{B}_{n,j}\varphi}\varphi$ and $\neg B_i^{\mathbb{B}_{n,j}\varphi}\varphi$ otherwise. This syntactic rendering is only intended to fix intuitions, though, since in countable models the limit of this sequence exceeds the finitary character of $\mathcal{L}_{EDL}$. We thus focus on model-theoretic conditioning.

Conditioning on a given event $A \subseteq W$ boils down to refining an agent's information partition by removing 'epistemic links' connecting A and non-A states.

Definition 4.16 (Conditioning by a subset)**.** *Given an epistemic plausibility model $\mathcal{M}$, the collection of epistemic equivalence relation of the agents is an element of $\wp(W \times W)^I$. Given a group $G \subseteq I$, the function $f_G : \wp(W) \to (\wp(W \times W)^I \to \wp(W \times W)^I)$ is a* conditioning function for G *whenever:*

$$(w, v) \in f_G(A)(i)(\{\sim_i\}_{i\in I}) \text{ iff}$$

$$\begin{cases} (w, v) \in \sim_i \textit{ and } (w \in A \textit{ iff } v \in A) & \textit{if } i \in G \\ (w, v) \in \sim_i & \textit{otherwise} \end{cases}$$

Given a model $\mathcal{M} = \langle W, (\leq_i)_{i\in I}, (\sim_i)_{i\in I}, V\rangle$ we write $f_G(A)(\mathcal{M})$ for the model $\langle W, (\leq_i)_{i\in I}, f_G(A)((\sim_i)_{i\in I}), V\rangle$.

It is easy to see that the relations $\sim_i$ in $f_G(A)(\mathcal{M})$ are equivalence relations. Here we are interested in cases where the agents condition their beliefs upon learning in which *belief state* the others are.

Definition 4.17 (Belief states)**.** *Let $\mathcal{M}$ be an epistemic plausibility model and $A \subseteq W$; we write*

$$B_j^{\mathcal{M}}(A) \textit{ for } \{w : \beta_j(\mathcal{K}_j^{\mathcal{M}}[w]) \subseteq A\} \textit{ and}$$

$$\neg B_j^{\mathcal{M}}(A) \textit{ for } W \setminus B_j^{\mathcal{M}}(A)$$

We define $\mathbb{B}_j^{\mathcal{M},w}(A)$ as follows:

$$\mathbb{B}_j^{\mathcal{M},w}(A) = \begin{cases} B_j^{\mathcal{M}}(A) & \textit{if } w \in B_j^{\mathcal{M}}(A) \\ \neg B_j^{\mathcal{M}}(A) & \textit{otherwise} \end{cases}$$

Observation 4.18. *For any plausibility epistemic model $\mathcal{M}$ indexed by a finite set of agents I, $\langle \wp(W \times W)^I, \subseteq \rangle$ is a chain complete poset. Moreover for all $A \subseteq W$, $w \in W$ and $G \subseteq I$, $f_G(A)$ is deflationary.*

Proof. Taking $\wp(W \times W)^I$ as a product, it is easy to see that $\langle \wp(W \times W)^I, \subseteq \rangle$ is a poset. The intersection of a decreasing sequence is the greatest lower bound of this sequence. Finally it is easy to see by inspecting Definition 4.16 that $f_G(A)$ is deflationary. Indeed for every i and A we have $f_G(A)(\sim_i) \subseteq \sim_i$ and thus by definition of a product $f_G(A)(\times_{i \in I}(\sim_i)) \subseteq \times_{i \in I}(\sim_i)$. QED

Taking $f_I(\bigcap_{j \in I} \mathbb{B}_j^{\mathcal{M},w}(||\varphi||^{\mathcal{M}}))$ as a mapping on models, it is easy to see from the preceding observation that conditioning by agents' beliefs about some event is deflationary with respect to the relation of epistemic-submodel. It follows then by the Bourbaki-Witt fixed point theorem [50] that conditioning by agents' beliefs has a fixed point.

Theorem 4.19 (Bourbaki-Witt [50])**.** *Let X be a chain complete poset. If $f : X \to X$ is inflationary (deflationary), then f has a fixed point.*

Given an initial pointed model $\mathcal{M}, w$ and some event $A \subseteq W$, we can construct its fixed point under conditioning by agents' beliefs as the limit of a sequence of models, which are the model-theoretic counterpart of the dialogues described above.

Definition 4.20. *A* conditioning dialogue about φ *at the pointed plausibility epistemic model $\mathcal{M}, w$, with $\mathcal{M} = \langle W, (\leq_i)_{i \in I}, (\sim_i)_{i \in I}, V \rangle$ is the sequence of pointed epistemic plausibility models $(\mathcal{M}_n, w)$ with*

$$(\mathcal{M}_0, w) = \mathcal{M}, w$$

$$(\mathcal{M}_{\beta+1}, w) = f_I(\bigcap_{j \in I} \mathbb{B}_j^{\mathcal{M}_\beta, w}(||\varphi||^{\mathcal{M}}))(\mathcal{M}_\beta), w$$

$$(\mathcal{M}_\lambda, w) = \bigcap_{\beta < \lambda} (\mathcal{M}_\beta, w) \text{ for limit ordinals } \lambda$$

This extends to the countable case the standard representation of a dialogue about φ in the literature on dynamic agreements [82, 14]. By observation 4.18 we know that dialogues cannot last forever, i.e. that each such sequence has a limit.

Corollary 4.21. *For any pointed epistemic plausibility model $\mathcal{M}, w$ and $\varphi \in \mathcal{L}_{EDL}$ there is an α^f such that, for all $i \in I$, $w \in W$ and $\alpha > \alpha^f$, $\mathcal{K}_{\alpha,i}[w] = \mathcal{K}_{\alpha^f,i}[w]$.*

Once the agents have reached this fixed point α^f—possibly after transfinitely many steps—they have eliminated all higher-order uncertainties concerning the posteriors about φ of the others, viz. these posteriors are then common knowledge:

Theorem 4.22 (Common knowledge of beliefs about φ)**.** *At the fixed point α^f of a conditioning dialogue about φ we have that for all $w \in W$ and $i \in I$, if $w \in B_i^{\mathcal{M}_{\alpha^f},w}(||\varphi||^{\mathcal{M}})$ then $w' \in B_i^{\mathcal{M}_{\alpha^f},w}(||\varphi||^{\mathcal{M}})$ for all $w' \in [w]^*_{\alpha^f,I}$, and similarly if $w \notin B_i^{\mathcal{M}_{\alpha^f},w}(||\varphi||^{\mathcal{M}})$.*

To save on notation we write $B_i^{\alpha^f}(A)$ for $B_i^{\mathcal{M}_{\alpha,f}}(A)$.

Proof. Let α^f be the fixed point existing by Corollary 4.21. Given an arbitrary state w in the domain of $\mathcal{M}_{\alpha^f}$ we prove that for any $w' \in [w]^*_{\alpha^f,I}$ and for any $i \in I$ we have $w \in B_i^{\alpha^f}(||\varphi||)$ iff $w' \in B_i^{\alpha^f}(||\varphi||)$. The proof is by induction on length of the smallest chain $C = \langle w_1 \sim_{\alpha^f,x} \cdots \sim_{\alpha^f,y} w_n\rangle$ where $x, y \in I$, $w_1 = w$ and $w_n = w'$.

Base case. For $|C| = 1$ is immediate by definition.

Induction step. We have two cases.

Case 1. $w \in B_{\alpha^f,i}(||\varphi||)$. Assume that we have a chain $C = \langle w_1 \sim_{\alpha^f,x} \cdots \sim_{\alpha^f,y} w_{n+1}\rangle$ where $x, y \in I$, $w_1 = w$ and $w_{n+1} = w'$ of length $n+1$. By IH we have $w_n \in B_{\alpha^f,i}(||\varphi||)$ (1). We have now two subcases. Subcase 1a: $w_n \sim_{\alpha^f,i} w'$ but then by epistemic introspection of beliefs and (1) we have $w' \in B_{\alpha^f,i}(||\varphi||)$. Subcase 1b: $w_n \sim_{\alpha^f,j} w'$ in C, for some $j \neq i$ in I. Now assume for a contradiction that $w' \notin B_{\alpha^f,i}(||\varphi||)$. It follows then by definition of the conditioning function $f_I(\bigcap_{j\in I} \mathbb{B}_j^{\mathcal{M},w}(||\varphi||^{\mathcal{M}}))$ that $f_I(\bigcap_{j\in I} \mathbb{B}_j^{\mathcal{M},w}(||\varphi||^{\mathcal{M}}))(\sim_{\alpha^f,j}) \subsetneq \sim_{\alpha^f,j}$. This contradicts the choice of α^f.

Case 2. The argument for the case of $w \notin B_{\alpha^f,i}(||\varphi||)$ is entirely similar, except that we use negative introspection of beliefs (if $w \notin B_i(X)$ then $\mathcal{K}_i[w] \subseteq \neg B_i(X)$). QED

With this in hand we can directly apply the static agreement result for common knowledge (Corollary 4.8, Section 4.3) to find that the agents do indeed reach agreements at the fixed point of a dialogue about φ.

Corollary 4.23 (Agreement via conditioning dialogue)**.** *Take any dialogue about φ with common and well-founded priors, and α^f as in Corollary 4.21. Then for all w in W, either $[w]^*_{\alpha^f,I} \subseteq \bigcap_{i\in I} B_i^{\mathcal{M}_{\alpha^f},w}(||\varphi||^{\mathcal{M}})$ or $[w]^*_{\alpha^f,I} \subseteq \bigcap_{i\in I} \neg B_i^{\mathcal{M}_{\alpha^f},w}(||\varphi||^{\mathcal{M}})$.*

This result brings qualitative dynamic agreement results [57, 14] to epistemic plausibility models, and shows that agents can indeed reach agreement via iterated conditioning, even when the finite model assumption is dropped.

4.5.2 Agreement via public announcements

In this section we show that iterated 'public announcements' lead to agreements, thus introducing a distinct form of information update to the agreement literature. We remind the reader that public announcements are 'epistemic actions' [67] by

which truthful, hard information is made public to the members of a group by a trusted source, in such a way that no member is in doubt about whether the others received the same piece of information as she did.

One extends a given logical language with public announcements by operators of the form $[\varphi!]\psi$, meaning "after the announcement of φ, ψ holds" [134, 83]. A dialogue about φ via public announcements among the members of a group G thus starts, as before, with i's simple beliefs about φ, for all $i \in I$. The agents' beliefs about φ at the next stage are then defined as those they would have after a public announcement of all agents' beliefs about φ at the first stage. Syntactically, this gives: $\mathbb{B}_{1,i}$ as in Section 4.5.1, and $\mathbb{B}_{n+1,i}$, as $[\bigcap_{j\in I} \mathbb{B}_{n,j}\varphi!]B_i\varphi$ if $\mathcal{M}, w \Vdash [\bigcap_{j\in I} \mathbb{B}_{n,j}\varphi!]B_i\varphi$ and as $[\bigcap_{j\in I} \mathbb{B}_{n,j}\varphi!]\neg B_i\varphi$ otherwise. For the same reason as in the previous section, we now move our analysis to the level of models.

Intuitively, the A-generated submodel (Definition B.4) of a given epistemic plausibility model $\mathcal{M}$ is the model that results after the public announcement of some formula φ, with $||\varphi||^{\mathcal{M}} = A$ in $\mathcal{M}$.

Definition 4.24 (Relativization by agents beliefs)**.** *Let $\mathbb{B}_i(\mathcal{M}, w, \varphi)$ be defined as follows:*

$$\mathbb{B}_i(\mathcal{M}, w, \varphi) = \begin{cases} ||B_i\varphi||^{\mathcal{M}} & \text{if } \mathcal{M}, w \Vdash B_i\varphi \\ ||\neg B_i\varphi||^{\mathcal{M}} & \text{otherwise} \end{cases}$$

Then given an epistemic-plausibility model $\mathcal{M} = \langle W, (\leq_i)_{i\in I}, (\sim_i)_{i\in I}, V\rangle$, the relativization $!B^{\varphi}_{w}$ by agents' beliefs about φ at w (where $w \in |\mathcal{M}|$), takes $\mathcal{M}$ to $!B^{\varphi}_{w}(\mathcal{M})$. Here $!B^{\varphi}_{w}(\mathcal{M})$ is the $\bigcap_{i\in I} \mathbb{B}_i(\mathcal{M}, w, \varphi)$-generated submodel $!B^{\varphi}_{w}(\mathcal{M}) = \langle W^{!B^{\varphi}_{w}}, \leq_i^{!B^{\varphi}_{w}}, \sim_i^{!B^{\varphi}_{w}}, V^{!B^{\varphi}_{w}}\rangle$ of $\mathcal{M}$ such that:

- $W^{!B^{\varphi}_{w}} = \bigcap_{i\in I} \mathbb{B}_i(\mathcal{M}, w, \varphi)$

and for each $i \in I$

- $\leq_i^{!B^{\varphi}_{w}} = \leq_i \cap\ (W^{!B^{\varphi}_{w}} \times W^{!B^{\varphi}_{w}})$
- $\sim_i^{!B^{\varphi}_{w}} = \sim_i \cap\ (W^{!B^{\varphi}_{w}} \times W^{!B^{\varphi}_{w}})$
- *For each $v \in W^{!B^{\varphi}_{w}}$, $v \in V^{!B^{\varphi}}(p)$ iff $v \in V(p)$*

Note that by the construction above the actual state w is never eliminated.

Observation 4.25. *For any plausibility epistemic model $\mathcal{M}$ indexed by a finite set of agents I, $\langle Sub(\mathcal{M}), \sqsubseteq\rangle$ is a chain complete poset. Moreover, for all $\varphi \in \mathcal{L}_{EDL}$, $w \in W$, $!B^{\varphi}$ is deflationary.*

Proof. It is easy to see that $\langle Sub(\mathcal{M}), \sqsubseteq\rangle$ is a poset. Moreover taking the submodel of $\mathcal{M}$ generated by the intersection of the domain of each element in a decreasing sequence is the greatest lower bound of this sequence. Finally it is easy to see by inspection of Definition 4.24 that $!B^{\varphi}$ is deflationary. QED

It follows then by the Bourbaki-Witt [50] Theorem (see previous subsection) that the process of public announcement of beliefs has a fixed point. Given an initial pointed model $\mathcal{M}, w$ and some formula $\varphi \in \mathcal{L}_{EDL}$, we can construct this fixed point by taking the limit of a sequence of models, which we call a public dialogue.

Definition 4.26. *A* public dialogue about φ *starting in* $\mathcal{M}, w$ *is a sequence of epistemic-doxastic pointed models* $\{(\mathcal{M}_n, w)\}$ *such that:*

- $\mathcal{M}_0 = \mathcal{M}$ *is a given epistemic-plausibility model.*
- $\mathcal{M}_{\beta+1} = !B^{\varphi}_{w}(\mathcal{M}_\beta)$
- $(\mathcal{M}_\lambda)$ *is the submodel of* $\mathcal{M}$ *generated by* $\bigcap_{\beta<\lambda} |\mathcal{M}_\beta|$ *for limit ordinals* λ

It is known that such a dialogue need not stop after the first round of announcements, in e.g. the 'muddy children' case [37], but by observation 4.25 we know that it will stop at some point.

Corollary 4.27 (Fixed point)**.** *Given an epistemic-plausibility model* $\mathcal{M}_0, w$ *and a public dialogue about* φ*, there is an* α^{φ} *such that* $(\mathcal{M}_\alpha, w) = (\mathcal{M}_{\alpha^{\varphi}}, w)$ *for all* $\alpha \geq \alpha^{\varphi}$*.*

Moreover at $\mathcal{M}_{\alpha^{\varphi}}, w$, which we call the *fixed point* of the public dialogue about φ, the posteriors of the agents about this formula are common knowledge, which means that they will reach an agreement on φ if they have common and well-founded priors.

Theorem 4.28 (Common knowledge at the fixed point)**.** *At the fixed point of a public dialogue* $\mathcal{M}_{\alpha^{\varphi}}, w$ *about* φ*, for all* $w \in W$ *and* $i \in I$*, if* $w \in ||B_i\varphi||^{\mathcal{M}_{\alpha^{\varphi}}}$ *then* $w' \in ||B_i\varphi||^{\mathcal{M}_{\alpha^{\varphi}}}$ *for all* $w' \in [w]^{*}_{\alpha^{\varphi},I}$*, and similarly if* $w \notin ||B_i\varphi||^{\mathcal{M}_{\alpha^{\varphi}}}$*.*

Proof. The proof follows the same line as for Theorem 4.22. QED

Corollary 4.29 (Agreements via Public Announcements)**.** *For any public dialogue about* φ*, if there are common and well-founded priors then at the fixed point* $\mathcal{M}_{\alpha^{\varphi}}, w$ *either all agents believe that* φ *or they all do not believe that* φ*.*

This new form of dynamic agreements result is conceptually important because it fits better than iterated conditioning the intuitive idea of a *public* dialogue, or so we shall argue in the next section, by highlighting the differences between the two processes of information exchange.

4.5.3 Comparing agreement via conditioning and public announcements

In this section we highlight by way of two examples that public announcements, in comparison with belief conditioning, are indeed *public*. We illustrate this first by comparing how conditioning and public announcements respectively change higher-order information, even in the case of 'non-epistemic' facts. We then point out that this difference can indeed lead to different agreements, precisely in cases where the dialogues *are* about epistemic facts.

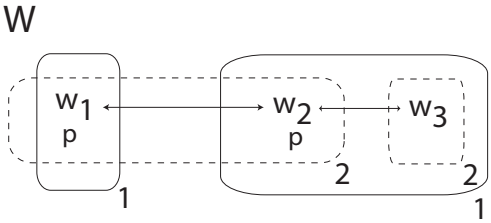

Figure 4.6: An epistemic plausibility model where one round of conditioning on p does not remove higher-order uncertainty about p, while a public announcement of p does.

Example 4.30. *Consider the model in Figure 4.6. The arrows represent* 1 *and* 2*'s common plausibility ordering, with* $w \leq w'$ *and* $w' \leq w$ *for all* $w, w' \in W$*. The solid and dotted rectangles represent* 1 *and* 2*'s information partitions, respectively. Take a proposition letter* p *and assume that* $V(p) = \{w_1, w_2\}$*. Observe that the agents already agree on* p *at* w_1*, but that agent* 2 *is uncertain about* 1*'s beliefs about* p*: writing* $\Diamond_2\psi$ *for* $\neg B_2\neg\psi$*, we have* $w_1 \models \Diamond_2 B_1 p \wedge \Diamond_2 \neg B_1 p$*. A single public announcement of* p *at* w_1 *suffices to remove this higher-order uncertainty:* $w_1 \models [p!]C_{\{1,2\}}p$*. Agent* 2*'s uncertainty about* 1*'s beliefs about* p*, however, remains after a single conditioning on* p*. Taking* $\Diamond_2^{\psi}\varphi'$ *for* $\neg B_2^{\psi}\neg\psi'$*, we have* $w_1 \models \Diamond_2^{p} B_1 p \wedge \Diamond_2^{p}\neg B_1 p$*.*

This example illustrates the public character of announcements in comparison with the private character of conditioning. In the first case all agents know that all others have received the same piece of truthful information. This is not necessarily the case for conditioning, even if all agents condition simultaneously on the same piece of information.

Given any pointed epistemic plausibility model $\mathcal{M}, w$ and formula φ, the reader can check that both the dialogue about φ via public announcements and the dialogue about φ via belief conditioning at $\mathcal{M}, w$ lead to the same agreement whenever φ is a Boolean combination of propositional letters. This is mainly due to the fact that neither operation changes the 'basic facts', i.e. the propositional valuation in a given model. They do, however, treat 'informational' facts differently, as the following example shows.

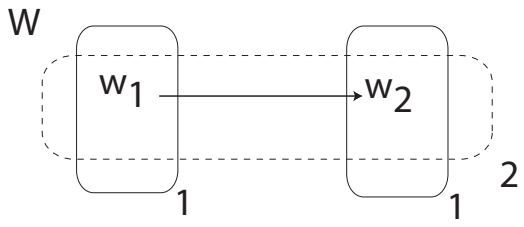

Figure 4.7: An epistemic plausibility model where conditioning leads to a different agreement than public announcements.

Example 4.31. *Consider the epistemic plausibility model in Figure 4.7. The arrows and rectangles are as in example 4.30. Take a proposition letter p and assume that* $V(p) = \{w_1\}$. *Let* $\varphi := p \wedge \neg B_2 p$, *i.e. "p but* 2 *doesn't believes that p". Observe that* φ *holds at* w_1, *that* 1 *believes it but that* 2 *does not. The conditioning dialogue and the dialogue via public announcements, both about* φ, *reach their fixed point* n^* *after one round in this model, where* $[w_1]_{n^*,1} = [w_1]_{n^*,2} = \{w_1\}$. *The formula* φ *leads to an 'unsuccessful update' by public announcement [67], and at the fixed point of the dialogue neither* 1 *nor* 2 *believe that* φ. *In conditioning dialogue, however, both agents* do *believe that* φ *at the fixed point.*

This example hinges on the fact that public announcement and belief conditioning have a different influence on higher-order information. In conditioning the truth value of the formula under consideration remains fixed. If the formula contains epistemic (K_i or C_G) or doxastic (B_i, CB_G) operators, this means that the conditioning dialogue bears on the knowledge and beliefs of the agents anterior to the information exchange [17]. In dialogues via public announcements the truth value of the formula φ is dynamically adapted to the incoming new information, reflecting the fact that knowing that others receive the same piece of information might lead an agent to revise his higher-order information, too.

This highlights the public character of announcements in comparison with belief conditioning, and thus that the former fit well with the intuition of public dialogue that drives the dynamic agreement results.

4.6 Definability of fixed points

We have seen that the static agreement result (Corollary 4.8) had a syntactic counterpart in some (undecidable, non-recursively enumerable) hybrid logic. What about the dynamic agreement results? Let us focus on the dynamic agreement results through public announcements. Here some remarks made in van Benthem [31] can guide us.

[31]'s 'rational dynamics' model steps of reasoning of agents in epistemic models for games. Here we do not focus so much on this issue (which we get back

to in Chapter 6), but rather on the technical issue raised about the *definability of fixed points*. Given some epistemic plausibility model $\mathcal{M}$ and some doxastic formula φ one can repeatedly announce φ until a fixed point is reached. We call this fixed point the φ-announcement limit of $\mathcal{M}$. But a public announcement need not be monotonic. Indeed it is possible that $\mathcal{M}$ is a submodel of $\mathcal{M}'$ but that $\mathcal{M}|\varphi$ is not a submodel of $\mathcal{M}'|\varphi$. But existential formulas are an interesting case. Indeed van Benthem [31] shows that:

Theorem 4.32 (van Benthem [31])**.** *The public announcement of φ is monotonic for existential doxastic epistemic formulas. As a corollary the φ-announcement limit is definable in the doxastic epistemic μ-calculus.*

What about announcement of beliefs? The bad news is that the announcement of the beliefs of the agents about some proposition need not be monotonic.

Fact 4.33. *Belief announcement is not monotonic.*

Proof. To see this the reader can look at Figure 4.8. Let the unfilled states be p-states. Assume we are in $\mathcal{M}'$, in the leftmost information partition: announcing the agent's belief will remove the rightmost state in the second model but be harmless in the first one. Clearly the resulting models are incomparable — none of them is a submodel of the other. If we are in the rightmost state things are even clearer: the resulting model after belief announcement in the bottom model is in fact a submodel of the one generated from the top one. QED

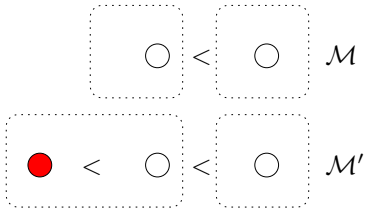

Figure 4.8: Belief announcement is not monotonic.

An immediate consequence of this fact (see Dawar et al. [59], van Benthem [31]) is that belief announcement won't be definable in the doxastic epistemic μ-calculus.

A syntactic confirmation of this bad news but also an interesting feature of the announcement scenario is that announcing 'disagreement' is really announcing the conjunction of a positive and a negative formula: 'one agent believes that φ, another does not believe that φ', thus if one conjunct is a positive formula the other one is bound to be a negative formula.

What about definability in more complex logics such as inflationary fixed point modal logics? In general inflationary fixed point modal logics will do for any fixed formula.

Theorem 4.34 (van Benthem [31])**.** *The limit submodel of iterated announcement of any doxastic epistemic formula is definable in the inflationary μ-calculus.*

First note that disagreement about φ is really a disjunction of two formulas: $(B_1\varphi \wedge \neg B_2\varphi) \vee (\neg B_1\varphi \wedge B_2\varphi)$. So we can define a map on models corresponding to announcement of 'disagreement about φ'. In which case the preceding result applies. So the limit of the announcement of disagreement about φ is definable in the inflationary μ-calculus.

But note that this is not exactly the type of announcement we were using in our earlier results. Indeed we required not only announcement of 'disagreement about φ' but of whether the agents believe or not that φ, i.e. of the particular disjunct that holds. It is unknown to us whether the weaker announcement of 'disagreement' guarantees common knowledge (or common belief) of posteriors in the inflationary fixed point under announcement of 'disagreement', hence whether agents will agree in the fixed point.

We could still have definability if the disagreement map was stable in the sense that if agents keep disagreeing they keep the same position, they don't switch their opinions. But even with a common prior this situation can happen!

Fact 4.35. *There exists a pointed model $\mathcal{M}, w$ satisfying common prior with $\mathcal{M}, w \Vdash B_1p \wedge \neg B_2p$ such that $\mathcal{M}|(B_1p \wedge \neg B_2p), w \Vdash \neg B_1p \wedge B_2p$*

Proof. Figure 4.9 presents such a situation.

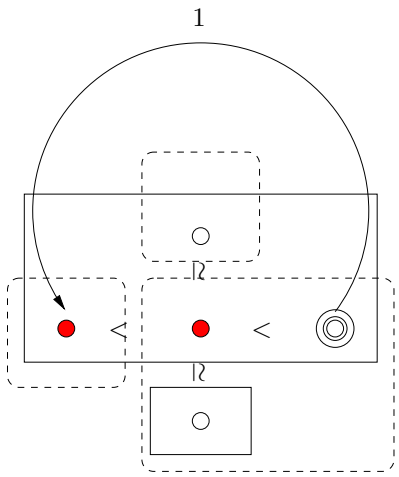

Figure 4.9: Agents can switch opinions. (Initial situation).

In the model in which 1 (plain information partition) and 2 (dashed information partition) have a common prior, red (darker) states are p-states and in the actual state (the double-circled, rightmost state), 1 believes that p while 2 does not. But let us announce their beliefs publicly...

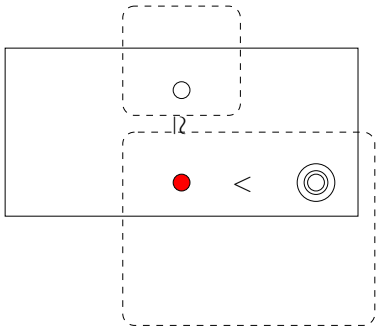

Figure 4.10: Agents can switch opinions. (Resulting situation).

... in the resulting model, the situation is now reversed: 2 now believes p while 1 does not believe p anymore. So it remains open whether the earlier results can be formulated within inflationary fixed points. QED

So we are really dealing with a conditional announcement of the form

$$(?\varphi_1; !\varphi_1) \cup (?\varphi_2; !\varphi_2)$$

It is unknown to us whether such conditional announcements can be defined in the inflationary μ-calculus. The same question can be raised for the more general case in which the announced formula is not necessarily the same as the one used as a precondition.

Public announcements of beliefs really represent a type of disagreement-solving scenario in which agents take belief announcements as hard information. In the preceding chapters we have extensively discussed soft dynamics such as lexicographic upgrade: what if agents instead of eliminating incompatible states, simply re-arrange their plausibility ordering, will we still get agreement in the limit? This is an open question. First, note that common knowledge of posteriors about φ is less demanding than common knowledge about the underlying proposition. The simple example in Figure 4.1 shows that in some cases common knowledge of posteriors might be reached via softer types of treatment of the other agent's belief (they will reach common knowledge of agreement in a single step, but obviously Enzo will never get to know that $\neg p$). But it still sounds more reasonable to hope for common *belief* of posteriors, which we learnt is sufficient to guarantee agreement.

About the issue of long term behavior of beliefs and group beliefs under soft announcements the reader can consult the recent work of Baltag and Smets [18]. How these results can be applied to the study of dynamic agreement results based on soft updates is an interesting open question.

On the syntactic side the issues we raised about definability of limits of belief announcements can be raised again for lexicographic upgrade. But now the

definability of limits of lexicographic upgrades will be more delicate. Indeed lexicographic upgrade defines a new (binary) relation, while public announcement simply takes a submodel. We can no longer expect definability in the inflationary μ-calculus, but rather in the fixed point extension of first-order logic: LFP(FO). The exact fragment of LFP(FO) in which dynamic agreement arguments based on lexicographic upgrade (and soft updates in general) can be expressed, is unknown to us. Determining it would be a source of great insight. Let us stress that similar technical issues are encountered in Zvesper [152, ch.3].

To conclude let us put the disagreement scenarios in a larger picture. Like the Muddy Children (but unlike the announcements of rationality of van Benthem [31]) the announcements in scenarios of disagreement-solving via public announcement are self-refuting in the limit (provided that the agents have a common prior). But unlike the ones in the Muddy Children they are not monotonic. But there are more scenarios describing how agents can communicate in order to try to solve their disagreement. In general it would be interesting to study the effect of various types of protocol restricting how agents are allowed to communicate what they believe. In general connections might exist with other formal or informal models of dialogical conflict resolution.

On a more technical side, it remains open whether one can finitely axiomatize a logic which can derive the static agreement results. Since the mere definability of fixed points is also open, so is the possibility of a finitary syntactic derivation of the dynamic agreement results. The expressibility of alternative agreement results, as e.g. the one provided in [142] is also open.

4.7 Conclusion

We have studied agreement theorems from the point of view of dynamic-epistemic logic. We have shown that both static and dynamic agreement results hold in epistemic plausibility models, answering an open question in the logic literature. Furthermore we have discussed syntactic counterparts for these results.

Major sources. The starting point of this chapter is naturally the work done on agreement theorems in the interactive epistemology literature: in particular Aumann [13]'s agreement theorem, Geanakoplos and Polemarchakis [82]'s dynamic results proving convergence via dialogues and the qualitative agreement results of Cave [57], Bacharach [14]. Another source is the sequence of papers by Baltag and Smets [16, 17] developing epistemic plausibility models and matching languages. The Σ^1_1-hardness of satisfiability for the hybrid logic considered in the syntactic derivation of the agreement result over transitive well-founded frames is a result from ten Cate [55]. Our discussion on definability of fixed points in modal languages is inspired by the one developed in van Benthem [31].

Our main results. In this chapter we considered the agreement results of interactive epistemology from a logical viewpoint. We proved static and dynamic

agreement results for epistemic plausibility models, reinforcing qualitative agreement results by proving that common *belief* in posteriors is sufficient to ensure agreement, under common and well-founded priors, and by proving so for both finite and countable structures. We pointed out the need for rather expressive logical languages to reason explicitly about static agreement results and gave a syntactic derivation of one of our static agreement results. Finally, we focused on the distinction between conditioning and public announcements to provide two dynamic agreement results.

The next step. This concludes our introduction of agreement theorems into the dynamic-epistemic logic area, bridging interactive epistemology and dynamic logics of belief and information. After interactive reasoning, our next step on the agenda is to consider inductive reasoning, and bring our logical perspective to the conditions under which an agent can reliably converge to some conjecture about its environment.

Chapter 5

Learning from the perspective of modal logics of belief change[1]

We have encountered doxastic epistemic temporal logics and dynamic doxastic logics and clarified their connection. Moreover we have seen how they can be used to analyze agreement issues in qualitative structures. In this chapter we are interested in how these logics can help us analyze inductive reasoning. More precisely this chapter investigates the connection between formal learning theory and modal logics of belief change and builds bridges between the two frameworks, bringing a logical point of view to inductive reasoning.

5.1 Introduction

Formal learning theory [107] brings a formal perspective to epistemological issues raised by the study of inductive reasoning. It models agents as functions that identify a correct hypothesis from a range of possibilities on the basis of inductively given streams of data. These functions can be viewed as agents that change their beliefs about which hypothesis is correct as they receive new information, according to some protocol. It is therefore natural to explore if some modal logics of belief change can give the syntactic means of analyzing inductive learning, giving interesting insights into its semantics and the pattern of reasoning at work in its analysis. To do so, we examine the temporal doxastic structure underlying formal learning theory, and look at how important concepts can be defined in modal logics of belief change.

The conceptual and philosophical background of this connection comes from a sequence of papers by Gierasimczuk, notably [85]. In what follows we focus on the language learning paradigm, which analyses the conditions under which an agent converges to a (correct grammar for a) language given a space of possible languages (resp. grammars) and does so, on any possible enumeration of that

[1]This chapter is based on Dégremont and Gierasimczuk [61].

language, treating languages as sets of positive integers. We refer to [85] for philosophical discussion and now move on to outline the formal details of the bridging.

In formal learning theory (FLT), learning is viewed as a process in which an agent (Learner) considers some range of languages. One of the languages is the actual one, and Learner's aim is to get to know which one it is. Elements of the language are given to Learner one by one. The infinite sequence of data that governs this enumeration includes all and only elements of the language. Several success conditions for Learner can be defined. For instance, we can assume that each time Learner gets a piece of information, she can make a conjecture. We can define the learning process to be successful if Learner's conjectures stabilize on the proper language. This learnability condition is called *identification in the limit* [87]. A more restrictive notion requires that Learner gives an answer only once, at some finite stage of the procedure. This kind of learnability is known as *finite identification* [123]. In Section 5.2 one can find a formal account of identification in FLT.

Intuitively, our approach to inductive learning in the context of modal logics of belief change (presented in Section 5.3) is as follows. We take the initial class of languages to be states in an epistemic plausibility model, which mirrors Learner's initial uncertainty and preferences over the range of languages. Each state (language) is assigned a protocol that indicates which sequences of events it allows (which streams of data enumerate that language). The incoming piece of information is taken to be an event that modifies the initial model. The structure resulting from updating the model with a sequence of events generates a doxastic epistemic temporal forest. We formulate the translation in Section 5.4.1.

We build on this construction in two ways. Firstly, we give a modal characterization of forests generated from a learning situation that satisfies a given learning condition (Section 5.4.2). Abstracting from this construction, we consider learnability conditions as properties that doxastic epistemic temporal models may or may not satisfy and show how FLT characterization theorems have natural counterparts in representation theorems about temporal protocols (Section 5.4.3). Finally in Section 5.5 we make a few observations on how to extend this approach to model situations in which learner might have imperfect observational power and might thus be affected by the presence of other learners. Section 5.6 concludes and presents directions for further work.

5.2 Formal learning theory

Let us start with some background on formal learning theory. Since FLT studies the conditions under which an agent can reliably converge to some language given a space of possible languages on any possible enumeration of that language, there are three things we need to make explicit: what classes of languages we are

interested in, what we mean by an enumeration (or a positive presentation) of a language, and most importantly what it means for an agent to 'reliably converge' or 'identify' a language on a given class. In fact we will not consider one but three notions of identification. But let us start with the classes of languages.

We treat languages as recursively enumerable sets $S \subseteq \mathbb{N}$. We will also be concerned with classes of such languages $\Omega = \{S_1, S_2, \ldots\}$. The indices will serve as names of the sets, or in other words, hypotheses.

Before discussing notions of identification let us fix the idea of a positive presentation of a language (i.e. of a set).

Definition 5.1. *By a* positive presentation (text) *of S, ε, we mean an infinite sequence of elements from S such that it enumerates all and only the elements from S allowing repetitions.*

We are almost ready to introduce the notions of identification. We only need to fix some basic notation about sequences.

Definition 5.2 (Notation)**.** *We will use the following notation:*

- *$U = \bigcup \Omega$ is the universal set of Ω;*
- *ε_n is the n-th element of ε; $\varepsilon|n$ is the sequence $(\varepsilon_0, \varepsilon_1, \ldots, \varepsilon_{n-1})$;*
- *$set(\varepsilon)$ is the set of elements that occur in ε;*
- *L is a learning function — a partial map from finite data sequences to indexes of sets, $L : U^* \rightharpoonup \mathbb{N}$.*

We can now give formal definitions of what it means for an agent to 'reliably converge to' or 'identify' a language on a given class (and to identify a class of languages). The first notion, finite identifiability, is the most demanding one, since it requires the learner to give a correct answer after a finite number of steps of enumeration of positive data.

Definition 5.3 (Finite identification; Gold [87], Mukouchi [123])**.** *A learning function L:*

1. *finitely identifies $S_i \in \Omega$ on ε iff, when inductively given ε, at some point L outputs i, and stops;*
2. *finitely identifies $S_i \in \Omega$ iff it finitely identifies S_i on every ε for S_i;*
3. *finitely identifies Ω iff it finitely identifies every $S_i \in \Omega$.*
4. *Ω is finitely identifiable iff some learning function L finitely identifies Ω.*

Let us illustrate these definitions with two examples from [85].

Example 5.4. $\Omega_1 := \{S_i = \{0, i\} | i \in \mathbb{N}\}$. Ω_1 *is finitely identifiable by* $L : U^* \to \mathbb{N}$*:*

$$L(\varepsilon|n) = \begin{cases} \text{is undefined if } set(\varepsilon|n) = \{0\}, \\ \max(set(\varepsilon|n)) \text{ otherwise.} \end{cases}$$

In other words, L outputs the correct hypothesis as soon as it receives a number different than 0*, and the procedure ends.*

Example 5.5. *To see how restrictive the notion of finite identifiability is, take a finite class of finite languages* $\Omega_2 = \{S_1, S_2, S_3\}$*, where* $S_i = \{1, \ldots, i\}$*.* Ω_2 *is not finitely identifiable. To see that, assume that* S_2 *is the actual language. A learning function can never conclude that* S_2 *is the actual language. For all it knows,* 3 *might appear in the future, so it has to leave the* S_3*-possibility open.*

The condition of finite identifiability might thus be seen as too demanding. Now if we allow Learner to answer each time she gets a new piece of data, we can define success as convergence to the right answer. This leads to the notion of identification in the limit.

Definition 5.6 (Identification in the limit [87])**.** *A learning function L:*

1. *identifies* S_i *in the limit on* ε *iff for co-finitely many m,* $L(\varepsilon|m) = i$*;*
2. *identifies* S_i *in the limit iff it identifies* S_i *in the limit on every* ε *for* S_i*;*
3. *identifies* Ω *in the limit iff it identifies in the limit every* $S_i \in \Omega$*.*
4. Ω *is identifiable in the limit iff some learning function identifies* Ω *in the limit.*

The following sequence of examples shows the notion of identification in the limit at work in the analysis of concrete learning situations.

Example 5.7. *First let us consider an example of a finite class of finite sets. Recall the class* Ω_2 *from Example 5.5.* Ω_2 *is identifiable in the limit by the following function* $L : U^* \to \mathbb{N}$*:* $L(\varepsilon|n) = m$*,* such that $m = max(set(\varepsilon|n))$*.*

Example 5.8. *The learning function from Example 5.7 identifies in the limit the following infinite class of finite sets:* $\Omega_3 = \{S_i | i \in \mathbb{N} - \{0\}\}$*, where* $S_n = \{1, \ldots, n\}$*.*

Example 5.9. *Identifiability in the limit of the class* Ω_3 *is lost if we enrich it by the set of all natural numbers. Let* $\Omega_4 = \{S_i | i \in \mathbb{N}\}$*, where* $S_0 = \mathbb{N}$ *and for* $n \geq 1$*,* $S_n = \{1, \ldots, n\}$*.* Ω_4 *is not identifiable in the limit. To see this, assume that there is a function L that identifies* Ω_4*. Then, there is a k and n, such that for all* $m \geq n$*,* $L(\varepsilon|m) = k$*. Now, if* $k \in \{1, 2, 3, \ldots\}$*, then L cannot identify the set* $\mathbb{N}$*. On the other hand, if* $k = 0$ *then L cannot identify* $S_{max(set(\varepsilon|n))}$*. So, we get a contradiction, L cannot identify* Ω_4*.*

Another epistemically plausible way to learn is by the elimination of hypotheses that are implausible, e.g. hypotheses that are inconsistent with the incoming data. This paradigm is formalized in the framework of learning by erasing.

Definition 5.10 (Function stabilization)**.** *In learning by erasing we say that a function stabilizes to number k on environment ε iff for co-finitely many $n \in \mathbb{N}$: $k = \min\{\mathbb{N} - \{L(\varepsilon|1), \ldots, L(\varepsilon|n)\}\}$.*

Definition 5.11 (Learning by erasing [115])**.** *A learning function L:*

1. *learns $S_i \in \Omega$ by erasing on ε iff L stabilizes to i on ε;*
2. *learns $S_i \in \Omega$ by erasing iff it learns by erasing S_i from every ε for S_i;*
3. *learns Ω by erasing iff it learns by erasing every $S_i \in \Omega$.*
4. *Ω is learnable by erasing iff some learning function learns Ω by erasing.*

It is easy to observe that in this setting learnability heavily depends on the chosen enumeration of languages, since the positive conjecture of the learning function is interpreted as the minimal one that has not been eliminated yet.

Now that the important notions behind the approaches of formal learning theory to inductive reasoning have been clarified, let us move to our logical approach.

5.3 Modal logics of belief change

Our logical approach to inductive learning will really be living at the interface of temporal and dynamic logics of belief change we have encountered in the first chapters, so we only need a few new ideas and concepts before we can start presenting its details.

5.3.1 Temporal models and languages for belief change

The reader will recall that doxastic epistemic temporal logics offer a global view of the evolution of a multi-agent system as events take place, focusing on the information that agents possess and what they believe. We gave a few variations on epistemic temporal models in Section 1.6.2. These variations apply to their doxastic epistemic cousins in the obvious way. This being said, in this chapter, we will be interested in logics interpreted on ω-W-doxastic epistemic temporal forests (cf. Section 1.6.2). Concerning these models we will refer to two *assumptions about doxastic and epistemic agents* that we have not considered so far.

Definition 5.12. *Let $\mathcal{H} = \langle W, \Sigma, H, (\sim_j)_{j \in \mathcal{A}}, V \rangle$ be an epistemic temporal model.*

Perfect Observation $\mathcal{H}$ *satisfies perfect observation iff* $\forall he, h'f \in H$ *if* $K_i[he] = K_j[h'f]$, *then* $e = f$. Perfect observation is satisfied if agents always know exactly what is happening.

Preference Stability $\mathcal{H}$ *satisfies preference stability iff* $\forall he, h'f \in H$ *we have* $he \leq_i h'f$ *iff* $h \leq_i h'$. It states that agents do not change their mind about the *a priori* plausibility of two histories as events take place. Naturally, it does not mean that the posterior beliefs of the agents might not evolve. Indeed, beliefs are defined as the most plausible states *of an information partition* and the latter might change.

On the language side, the ideas behind the formalisms we will be using are already familiar to the reader.

A hybrid doxastic epistemic temporal language

We will be using the language $\mathcal{H}_{\text{BDET}}(\downarrow)$, i.e. the hybrid version of $\mathcal{L}_{\text{BDET}}$ (that we introduced in 3.3.2) with binders (cf. Section 1.7). This language will be interpreted over an ω-W doxastic epistemic temporal model $\mathcal{H}$, an initial state w, an infinite history $w\epsilon$ and a finite prefix wh of $w\epsilon$ together with an assignment function $g : \text{SVAR} \to H$, mapping state variables to nodes, i.e. state variables will be true in exactly one node (regardless of the infinite path under consideration). Explicitly we use the following additional clauses:

$$\begin{array}{lll} \mathcal{H}, w\epsilon, wh, g \Vdash x & \text{iff} & g(x) = wh \\ \mathcal{H}, w\epsilon, wh, g \Vdash \downarrow x.\varphi & \text{iff} & \mathcal{H}, w\epsilon, wh, g[x := wh] \Vdash \varphi \end{array}$$

And nothing more that the reader has not yet been exposed to on the temporal side.

5.3.2 The dynamic approach

As for the dynamic doxastic and dynamic epistemic logics perspective that considers belief change as stepwise operations on models, we will model, as previously, static doxastic epistemic situations as *epistemic plausibility models*. As for modeling the doxastic and epistemic update as events take place, we use epistemic event models (without plausibility orderings). Indeed this restricted perspective is all we need to capture the setting of finite identifiability. Other notions for identifiability might call for the richer setting of (epistemic) plausibility event models, but in this chapter we restrict ourselves to finite identifiability.

The effect of updating an epistemic plausibility model $\mathcal{M}$ by an event model $\mathcal{E}$ is computed according to (conservative) *product update* (Definition 2.17), which is intuitively a special case of Priority Update, in which the dynamics are purely

epistemic. Recall that an epistemic plausibility model describes what agents currently believe and know, while product update creates the new doxastic epistemic situation after some informational event has taken place.

Recall that a *protocol* P maps states in an epistemic plausibility model to sets of finite sequences of pointed event models closed under taking prefixes. This defines the admissible runs of some informational process. We let $\mathfrak{E}$ be the class of all pointed epistemic (plausibility) event models. Let $Prot(\mathfrak{E}) = \{\mathtt{P} \subseteq (\mathfrak{E}^* \cup \mathfrak{E}^\omega) \mid \mathtt{P} \text{ is closed under non-empty finite prefixes}\}$ be the co-domain of protocols, it is the class of all sets of sequences (infinite and finite) of pointed epistemic (plausibility) event models closed under taking finite prefixes.

Definition 5.13. *Let us take an epistemic plausibility model $\mathcal{M}$, and let $|\mathcal{M}|$ be the domain of $\mathcal{M}$. A local protocol for $\mathcal{M}$ is a function $P : |\mathcal{M}| \to Prot(\mathfrak{E})$.*

To refer to a doxastic epistemic model corresponding to the doxastic situation of the agent after she has received some information (e.g. after some positive enumeration of a language has started), we use the concept of a $P, \epsilon|n$-generated epistemic model.

Definition 5.14. *A $P, \epsilon|n$-generated* epistemic model $\mathcal{M}^{P,\epsilon|n}$ *is defined inductively in the following way: $\mathcal{M}^{P,\epsilon|0} = \mathcal{M}$; $\mathcal{M}^{P,\epsilon|n+1} = \langle |\mathcal{M}^{P,\epsilon|n+1}|, \sim_{P,\epsilon|n+1}, V_{P,\epsilon|n+1}\rangle$, where:*

1. $|\mathcal{M}^{P,\epsilon|n+1}| := \{s\epsilon|n+1 \mid s\epsilon|n \in \mathcal{M}^{P,\epsilon|n}$ *and* $\epsilon|n+1 \in P(s)\}$;
2. $\sim_{P,\epsilon|n+1} := \sim_{P,\epsilon|n} \cap (|\mathcal{M}^{P,\epsilon|n+1}| \times |\mathcal{M}^{P,\epsilon|n+1}|)$;
3. *For every* $p \in$ PROP, $V_{P,\epsilon|n+1}(p) := V_{P,\epsilon|n}(p) \cap |\mathcal{M}^{P,\epsilon|n+1}|$.

The connection between the two approaches has been discussed in depth in the two first chapters, so we simply state a definition we have not formally encountered so far (though the idea may now be fairly natural for the reader).

5.3.3 Connecting the temporal and the dynamic approach

In Section 2.5 we gave the formal definitions of the doxastic forest generated via a sequence of priority updates. In the particular case where we consider only *epistemic update* of epistemic plausibility models, the DETL forest generated can be obtained as follows:

Definition 5.15 (DETL forest generated by a state-dependent DEL-protocol)**.** *Each initial epistemic plausibility model $\mathcal{M} = \langle W, (\sim_i^{\mathcal{M}})_{i\in\mathcal{A}}, (\leq_i^{\mathcal{M}})_{i\in\mathcal{A}}, V^{\mathcal{M}}\rangle$ and each local protocol P yields a generated DETL forest $For(\mathcal{M}, P)$ of the form: $\mathcal{H} = \langle W^{\mathcal{H}}, \Sigma, H, (\sim_i)_{i\in\mathcal{A}}, (\leq_i)_{i\in\mathcal{A}}, V\rangle$, as follows:*

1. $W^{\mathcal{H}} = |\mathcal{M}|$, $\Sigma = \bigcup_{w\in W} \bigcup_{n\in\mathbb{N}} P(w)(n)$,

2. *H is defined inductively as follows: $H_1 = W^{\mathcal{H}}$;*

   - $H_{n+1} := \{(we_1 \dots e_{n+1}) \mid (we_1 \dots e_n) \in H_n;$
     $\mathcal{M} \otimes \epsilon_1 \otimes \dots \otimes \epsilon_n, (w, e_1, \dots, e_n) \Vdash \mathtt{pre}_n(e_{n+1})$ *and* $e_1 \dots e_{n+1} \in P(w)\}$*;*
   - $H = \bigcup_{1 \leq k < \omega} H_k$.

3. *If $h, h' \in W^{\mathcal{H}}$, then $h \sim_i h'$ iff $h \sim_i^{\mathcal{M}} h'$;*

4. *For $1 < k \leq m$, $he \sim_i h'e'$ iff $he, h'e' \in H_k$, $h \sim_i h'$, e and e' are events from the same event model and $e \sim_i e'$ in their event model;*

5. *For $1 < k \leq m$, $he \leq_i h'e'$ iff $he, h'e' \in H_k$ and $h \leq_i h'$;*

6. *Finally, $wh \in V(p)$ iff $w \in V^{\mathcal{M}}(p)$.*

It is easy to obtain a representation result for this type of generated forest as a corollary of Benthem et al. [36]'s Theorem 2.3 introduced in Section 2.2. To do so, it is sufficient to require that the corresponding doxastic epistemic temporal frames satisfy Preference Stability.

Corollary 5.16. *Let $\mathcal{H}$ be an arbitrary epistemic-temporal DETL model. The following two assertions are equivalent:*

- *$\mathcal{H}$ is isomorphic to the DETL forest generated by the sequential product update of some epistemic* plausibility *model according to some state-dependent DEL-protocol P*

- *$\mathcal{H}$ satisfies Propositional Stability, Synchronicity, Bisimulation Invariance, Perfect Recall, Uniform No Miracles* and Preference Stability.

This concludes this refresher and the exposition of the few ideas we didn't yet encounter at this stage. Let us now move to our logical analysis of inductive learning.

5.4 Analyzing learnability in a DETL framework

This section gives the first results bridging learning theory and dynamic epistemic temporal logics. We prove that the problem of checking whether a class of sets is finitely identifiable can be reduced to the model-checking problem of $\mathcal{H}_{\mathrm{BDET}}(\downarrow)$ on doxastic epistemic temporal forests. To start with, we show how learning situations can be encoded by an epistemic plausibility model and a local protocol.

5.4.1 Protocols that correspond to set learning

As we have seen learning scenarios involve a single learner so we take, $\mathcal{A} = \{L\}$ and write $\sim$ instead of $\sim_L$. Given a class of languages $\Omega = \{S_1, \ldots, S_i \ldots\}$ we define out initial epistemic model as follows:

Definition 5.17 (Initial epistemic model)**.** *Our initial epistemic model $\mathcal{M}_\Omega$ is a triple: $\langle W_\Omega, \sim_\Omega, V_\Omega\rangle$, where $W_\Omega = \Omega$, $\sim_\Omega = W_\Omega \times W_\Omega$. The valuation is irrelevant.*

In words, we identify states of the model with sets, we also assume that our agent does not have any particular initial information.

Definition 5.18 (Single event model)**.** *For each $e \in U$, we have a corresponding event model $\mathcal{E} = \langle\{e\}, \sim^{\mathcal{E}}, \mathtt{pre}_{\mathcal{E}}\rangle$ where $\sim^{\mathcal{E}} = \{(e, e)\}$ and $\mathtt{pre}_{\mathcal{E}}(e) = \top$.*

In words we assume that the agent knows exactly what event is happening (what element is being enumerated), i.e. she is a perfect observer. We now get to the interesting part of the construction.

Given a set S_i, we can transform it into a set of events, we write $\mathbb{E}(S_i) = \{(\mathcal{E}, e) \mid e \in S_i\}$. We trivialize the role of preconditions; the admissible sequences of events are defined by means of protocols.

We now define local protocols. Intuitively, given a state $S_i \in W_\Omega$, our protocol P_Ω should authorize at S_i any ω-sequence that enumerates S_i and nothing more.

Definition 5.19 (Local protocol)**.** *For every $S_i \in W_\Omega$, $P_\Omega(S_i)$ is the smallest subset of $(\mathbb{E}(U))^* \cup (\mathbb{E}(U))^\omega$ that contains $\{f : \omega \to \mathbb{E}(S_i) \mid f$ is surjective$\}$, and that is closed under non-empty finite prefixes.*

We have now identified the natural image of a learning situation within the structures living at the interface of temporal models and dynamic models of belief change. We now get to the logical analysis of finite identifiability.

5.4.2 DETL characterization of finite identifiability

We start by defining a DETL version of the notion of belief stabilization (resp. knowledge stabilization) to a certain hypothesis.

Definition 5.20. *An agent j's belief (resp. knowledge) about the initial state* stabilizes *to w on the history $v\epsilon$ iff there is a finite prefix $e^* \sqsubset \epsilon$ such that for any finite sequence e' such that $e^* \sqsubseteq e' \sqsubset \epsilon$ and for all histories sh such that $sh \in \mathcal{B}_j[ve']$ we have $s = w$ (resp. for $\mathcal{K}_j[ve']$).*

We can now show that checking whether an agent can reliably converge on some language within a class of languages after a finite number of steps of enumeration of positive data from this language (and so for all languages of that class), can be reduced to model-checking a formula of $\mathcal{H}_{\mathrm{BDET}}(\downarrow)$ in the forest generated from the preceding epistemic model by product updating it according to the preceding protocol.

Proposition 5.21. *The following are equivalent:*

1. *Ω is finitely identifiable.*
2. *In the generated forest $For(\mathcal{M}_\Omega, P_\Omega)$, for all $S_i \in W_\Omega$ and $\epsilon \in P_\Omega(S_i)$ the learner's knowledge about the initial state stabilizes to S_i on $S_i\epsilon$.*
3. *$For(\mathcal{M}_\Omega, P_\Omega) \Vdash \mathtt{A}(\neg\bigcirc^{-1}\top \rightarrow \downarrow x.\forall F\ KH(\neg\bigcirc^{-1}\top \rightarrow x))$.*

Proof. (1 $\Rightarrow$ 2) We prove the contrapositive. Assume that there is a state $S_i \in W_\Omega$ and ω-sequence $\epsilon \in P_\Omega(S_i)$ such that the agent's knowledge does not stabilize to S_i on ϵ. There are two cases.

1. The learner stabilizes to another state, but then by construction of $P_\Omega(S_i)$ and the definition of a generated DEL-forest for every finite prefix $h \sqsubset \epsilon$, $S_i h \in \mathcal{K}[S_i h]$. Contradiction. So we are in the other case.
2. After each finite prefix $h \sqsubset \epsilon$, there is at least one state different from S_i that remains epistemically possible. Since a DEL-generated ETL forest satisfies perfect recall (Theorem 2.3), it follows that there is some state $S_i \neq S_j$ that remains epistemically possible after each finite prefix $h \sqsubset \epsilon$. But by construction of $P_\Omega(S_i)$ this is only possible if $S_i \subset S_j$. Every finite subset of S_i is a subset of S_j, and therefore $S_i \in \Omega$ does not have a finite definite tell-tale set. Therefore, from Theorem 7 in [123], Ω is not finitely identifiable.

(2 $\Rightarrow$ 3) We prove the contrapositive. Assume that $For(\mathcal{M}_\Omega, P_\Omega) \nVdash \mathtt{A}(\neg\bigcirc^{-1}\top \rightarrow \downarrow x.\forall F KH(\neg\bigcirc^{-1}\top \rightarrow x)\ )$. This means that some history satisfies $\neg\bigcirc^{-1}\top$, i.e., there is some initial state in $w \in W_\Omega$, such that for some $\epsilon \in P_\Omega(w)$ and for every finite prefix $h \sqsubset \epsilon$ we have $For(\mathcal{M}_\Omega, P_\Omega)w, w\epsilon, wh, g[g(x) := w] \nVdash KH(\neg\bigcirc^{-1}\top \rightarrow x))$. By the truth conditions of K and $H(\neg\bigcirc^{-1}\top \rightarrow x))$, there is some history $vh' \in \mathcal{K}[wh]$ with $v \neq w$. But this means that Learner's knowledge does not stabilize to w on $w\epsilon$ in $For(\mathcal{M}_\Omega, P_\Omega)$. Contradiction.

(3 $\Rightarrow$ 1) By the semantics of $\mathtt{A}(\neg\bigcirc^{-1}\top \rightarrow$ we know that in every initial state $S_i \in W_\Omega$: $S_i \Vdash \downarrow x.\forall F KH(\neg\bigcirc^{-1}\top \rightarrow x)$ (1). Now assume for a contradiction that there is some S_i that is not finitely identifiable in Ω. It follows that there is some enumeration ϵ^* of the set such that after any finite prefix of ϵ^*, there is another set S_j that the agent has not excluded (2).

But by (1) we can label S_i by x and for any sequence of events ϵ, there will be a finite prefix $\epsilon|m$ at which $S_i\epsilon|m, \epsilon, g[x := S_i] \Vdash KH(\neg\bigcirc^{-1}\top \rightarrow x)$ (3). By construction of P_Ω we have a finite prefix $\epsilon^*|n$ such that $S_i\epsilon^*|n, \epsilon^*, g[x := S_i] \Vdash KH(\neg\bigcirc^{-1}\top \rightarrow x)$ (4). But then the agent knows that the initial state was $g(x) = S_i$ and thus has excluded any other initial state, contradicting (2). QED

The (1-3) equivalence shows that we can characterize finite identifiability by the (local or) global satisfaction of a formula from the hybrid doxastic epistemic

temporal language $\mathcal{H}_{\text{BDET}}(\downarrow)$. This is the part that shows that the problem of checking whether a class of sets is finitely identifiable can be reduced to the model-checking problem of $\mathcal{H}_{\text{BDET}}(\downarrow)$ on doxastic epistemic temporal forests. (1-2) equivalence indicates that we can abstract away from forests that are actually generated from learning situations and reason directly about DETL models. This is what we do in the next section.

5.4.3 Characterizing protocols that guarantee learnability

Indeed, recall the condition from Proposition 5.21:

> *In the generated forest* $For(\mathcal{M}_\Omega, P_\Omega)$, $S_i \in W_\Omega$ *and* $\epsilon \in P_\Omega(S_i)$ *the learner's knowledge about the initial state stabilizes to* S_i *on* $S_i\epsilon$.

We can abstract away from this condition to define a structural property that can be (or fail to be) satisfied by a given DETL frame. We start by generalizing finite identifiability.

Definition 5.22. *A DETL frame* $F(\mathcal{H}) = \langle W, \Sigma, H, \leq_L, \sim_L \rangle$ *satisfies finite identification (FIN) iff for all* $s \in W$ *and* $s\epsilon \in P(s)$ *Learner's* knowledge *about the initial state stabilizes to* s *on* $s\epsilon$.

We now define what it means for a DETL frame to satisfy the 'learning by erasing property'.

Definition 5.23. *A DETL frame* $F(\mathcal{H}) = \langle W, \Sigma, H, \leq_L, \sim_L \rangle$ *satisfies learning by erasing (ERASE) iff for all* $s \in W$ *and* $h = s\epsilon \in P(s)$ *Learner's* belief *about the initial state stabilizes to* s *on* $s\epsilon$.

We can now prove representation theorems that characterize classes of DETL frames in which learnability is guaranteed in terms of properties of the protocol the DETL model is based on. We start by giving two results about finite identification and then say a few words about a DETL counterpart of an important result of formal learning theory: Angluin's Theorem.

Proposition 5.24. *A synchronous, perfect recall, perfect observation DETL model* $\langle W, \Sigma, H, \sim, \leq, V \rangle$ *satisfies* finite identifiability *whenever for each* $w \in W$ *and history* $wh \in H \cap \Sigma^\omega$, *there is some natural number* $n \in \omega$ *such that for every* $v \neq w$ *such that* $v \in W$ *and for every* $vh' \in H \cap \Sigma^\omega$ *we have* $(h|n) \neq (h'|n)$.

Proof. Take an arbitrary w. By assumption there is some $n \in \omega$ such that for every $v \neq w$ such that $v \in W$ and for every $vh' \in H \cap \Sigma^\omega$ we have $(h|n) \neq (h'|n)$. We prove that $w(h|n) \not\sim v(h'|n)$ by induction. Indeed assume that they are in the same information partition. Then by perfect observation the last events were the same. But by perfect recall we also have that the nodes right before were also in the same information partition so we can iterate this argument and apply perfect observation all the way down, proving that $(h|n) \neq (h'|n)$. QED

The next result corresponds to the finite identifiability characterization [123].

Proposition 5.25. *A permutation closed, synchronous, perfect recall, perfect observation DETL model* $\langle W, \Sigma, H, \sim, \leq, V\rangle$ *based on a finite state space satisfies* finite identifiability *whenever for all* $w \in W$ *there is an event* $a \in \mathbb{E}(w)$ *such that for all* $v \in W$ *if* $v \neq w$*, then* $a \notin \mathbb{E}(v)$*.*

Sketch. Take an arbitrary $w \in W$. By hypothesis there is an event $a \in \mathbb{E}$ such that for each $v \neq w$ we have $a \notin \mathbb{E}(v)$. By permutation closure a is included in every $\epsilon \in P(w)$. But, by the definition of P we know that in every $\epsilon \in P(w)$ the event a occurs at some finite stage. Take an arbitrary $\epsilon \in P(w)$. For some $n \in \omega$, we have $\epsilon_n = a$. Now assume for a contradiction that at stage $n+1$ some state $v \neq w$ is still considered possible. But then it means that $a \in \mathbb{E}(v)$. Contradiction. QED

We now turn to a DETL counterpart to a crucial result in learning theory: Angluin's theorem, that characterizes classes of sets that are identifiable in the limit.

Theorem 5.26 (Angluin [5])**.** *A class of sets* Ω *is identifiable in the limit iff for all* $S \in \Omega$ *there is a finite* $D_S \subseteq S$ *such that for all* $S' \in \Omega$*, if* $S \neq S'$ *and* $D_S \in S'$*, then* $S' \not\subseteq S$*.*

The next result is proved in Dégremont and Gierasimczuk [61] using once more the concept of a DEL-generated forest. Before we state the result, referring to [61] for the proof, let us introduce the following definitions:

Set-driven A local protocol P for $\mathcal{M}$ is set-driven iff $\forall w \exists S_w \subseteq \mathbb{N}$ such that $\forall \varepsilon \in P(w)$ $set(\varepsilon) = S_w$.

A-condition for protocols A local protocol P satisfies the A-condition iff $\forall w \exists e \in P(w) \cap \Sigma^* \forall w \neq v (e \in P(v) \implies P(v) \not\subset P(w))$.

Finite identifiability of incomparable sets A local protocol P satisfies the condition of finite identifiability of incomparable sets iff states whose image under P are $\subseteq$-incomparable constitute finitely identifiable classes.

Let us assume that a local protocol P satisfies finite identifiability of the incomparable. Dégremont and Gierasimczuk [61] show the following equivalence.

Theorem 5.27 (Dégremont and Gierasimczuk [61])**.** *A state space* W *together with a set-driven local protocol* P *satisfies the A-condition iff there is a preference ordering* $\leq$ *on* W *and an epistemic plausibility frame* $M = (W, \sim, \leq)$*, where* $\sim = W \times W$ *such that*

(#) for all $w \in W$ and for all $\varepsilon \in P(w)$ there is some $n \in \omega$ such that for every $m > n$, $w \in |M^{\varepsilon|m}|$ and w is the $\leq$-minimum of $|M^{\varepsilon|m}|$ in the generated doxastic model $M^{\varepsilon|m}$.

This concludes our considerations on representations of DETL learnability properties in terms of properties of protocols. Before we conclude we would like to make a few observations about the possibility and interest of extending single-agent learning to multi-agent learning.

5.5 About multi-agent (interactive) learning

Clearly, learning is usually a multi-agent process. What changes if more agents are learning at the same time and are allowed to communicate? We start with an intuitive conjecture. Take a learning situation in which all agents have the same initial information and they are all perfect observers, i.e. there is no uncertainty for them as to which event is actually taking place. Now consider an arbitrary agent i: i will learn in the same way whether or not she can communicate with the other agents. So multi-agent learning is more interesting if one considers cases in which agents have imperfect observational power, i.e. if they might be uncertain about what Nature is "saying". So moving to the multi-agent case will be interesting if we drop the assumption that agents know exactly to what corresponds the signal they are observing. This would be equivalent to treating enumeration not as directly communicating positive information but as sending signals whose interpretation is not immediately transparent to the agents.

If we restrict the nature of this communication there are still cases in which learners are not affected by the presence of other learners. Indeed if agents have the same observational powers, whether or not we allow the learners to (truthfully) announce their current conjecture before each new item is enumerated, they will converge to the same hypothesis. Here is a modeling of this idea of 'conjecture announcement', that is slightly more abstract than in the previous chapter by seeing it as a particular event $!B$ having a special clause in the product update rule.

Definition 5.28 (Product Updating by $!B$)**.** *The* product update *of an epistemic model $\mathcal{M} = \langle W, (\sim_i)_{i\in N}, V\rangle$ by $!B$ is the model $\mathcal{M}\otimes !B$ with domain $W \times \{!B\}$ and whose epistemic relations are defined as:*

$$(w, !B) \sim_i' (w', !B) \text{ iff } \forall j \in N\ \mathcal{B}_j[w] = \mathcal{B}_j[w'], w \sim_i w'$$

and whose valuation is defined by

$$(w, !B) \in V(p) \text{ iff } w \in V(p)$$

Given a protocol P we refer to $P^{!B}$ as the result of interlacing belief announcement with each step of the inductive enumeration. We can now formally prove what we were claiming, namely that, with the same initial information and the same observational powers, considering the multi-agent dimension is irrelevant.

Theorem 5.29. *Let $\mathcal{M}$ be some epistemic plausibility model in which the agents have the same background information and let P be a protocol for $\mathcal{M}$ in which agents have the same observational powers. For each agent $j \in N$, states w in $|\mathcal{M}|$ and environment $\epsilon \in P(w)$, j's belief stabilizes to $v \in |\mathcal{M}|$ on $w\epsilon$ in the forest generated by $\mathcal{M}$ and P (without announcement of conjectures) iff it stabilizes to $v \in |\mathcal{M}|$ on $w(\epsilon)^{!B}$ in the forest generated by $\mathcal{M}$ and $P^{!B}$ (with announcement of conjectures).*

Proof. We prove this result in Appendix E. QED

Implicit assumptions about agents are naturally at work in the previous theorem due to the fact that we assume that the learning situation and learners' types were encoded within a *DEL*-type framework. Therefore by Theorem 2.3, we know that Theorem 5.29 tells us that if we are assuming our agents to satisfy Perfect Recall and Uniform No Miracles, then identical observational powers and initial information makes communication of conjectures redundant.

This absence of importance that learners assign to other beliefs is intuitively easy to understand: since they share the same information at every step of the process, differences of beliefs are only due to differences in prior and what agents announce is already expected by the agents. The preceding theorem is quite tight in the sense that Uniform No Miracles is the only assumption that can be weakened. The usual Muddy Children scenario (see e.g. [37, 67]) is an example where communication of beliefs *does* make a difference, when agents either don't have the same initial information or don't have the same observational capacities. For Perfect Recall the idea is that if an agent has only bounded memory, she might benefit from communicating with agents who have unbounded memory (and might remind her of what she forgets).

But from the previous chapter we have learned that agents might actually reach agreements. Indeed from Corollary 4.29 we can conclude that if agents have the same prior plausibility ordering and this prior plausibility ordering is well-founded, then they will eventually reach agreement — provided events (corresponding to the enumeration) map epistemic models to epistemic models. In particular if the uncertainty of agents about each other's observational capacities can be modeled by an epistemic event model, thus assuming that agents do not change the relative plausibility ordering of possible hypotheses, then we get the following:

Proposition 5.30. *Let $\mathcal{M}$ be some epistemic plausibility model in which the agents have the same well-founded prior and a protocol P for $\mathcal{M}$ in which agents*

have observational powers encoded by epistemic event models. If agents are allowed to keep publicly communicating their conjectures between each step until they possibly reach agreement, then they will eventually (possibly after transfinitely many steps between each new element enumerated) stabilize all to the same conjecture if they converge at all.

Proof. The idea of the proof is simple. Since product-updating an epistemic plausibility model by an epistemic event model keeps the relative ordering of conjectures unchanged, it also remains well-founded and remains the same between agents. Now by Corollary 4.29 we know that agents having the same well-founded prior announcing their beliefs will eventually reach agreement. It follows that after each new enumerated element (after each new signal sent) they will eventually re-reach agreement. So if one of them converges, they will all converge. QED

So assuming that agents are conservative with respect to the initial ordering of the possible hypotheses, and they have the same well-founded prior ordering, then communication of conjectures guarantees that agreement will be maintained and thus that if agents converge at all, they will all converge to the same conjecture.

From where we stand future work includes extending our approach to other types of identification, e.g., identification of functions (which generalizes the preceding setting in which sets are identified) or learning from both positive and negative information (rather than only positive information in the previous setting). Another line is to study the effects of different restrictions on protocols on identifiability. And similarly for various constraints on learning functions (e.g. consistency, conservatism or set-drivenness), comparing them to those of epistemic and doxastic agents in the DETL framework. Finally our modal characterization of finite identifiability carries with it a modal concept of 'stable belief' that can be extracted and studied for itself. In fact each notion of identifiability carries implicitly a notion of stable belief whose logic it would be interesting to axiomatize.

5.6 Conclusions and perspectives

We brought a logical perspective to inductive reasoning, studying identifiability in scenarios of set learning, at the interface of temporal and dynamic logics of belief change. We did so from both a more syntactic and a purely structural point of view. Finally we made a few observations about multi-agent (interactive) learning.

Major sources. The starting point of this chapter is the existing work on the connection between learning theory and dynamic epistemic logic by Gierasimczuk, notably [85]. For the set learning setting the source is Gold [87] and for the concepts of identifiability we discuss they are Gold [87], Mukouchi [123], Lange

et al. [115]. Sources of the logical framework are the ones mentioned in the first two chapters, in particular van Benthem et al. [36], and these first chapters themselves.

Our main results. This chapter has shown that the problem of finite identifiability of a class of sets can be reduced to the model checking problem of a hybrid branching-time epistemic temporal logic over DETL frames. First we have seen that the setting of a learning scenario could be represented in a DEL setting with protocols, and then we identified a formula that corresponds to finite identifiability. Abstracting on the notion of stabilization of belief (or knowledge) at work in this first result, we have defined an abstract concept of learnability for DETL frames and have given two representation theorems for such abstract conditions in terms of properties of the underlying protocol of the DETL frame. Finally we made a few observations about the extension of learning to multi-agent interactive scenarios in a more general setting where learners are not assumed to be perfect observers.

The next step. This concludes our logical study of inductive reasoning from the perspective given by the interface between temporal and dynamic doxastic logics, developing the connection of the latter with formal learning theory. After interactive reasoning and inductive reasoning, we would like to make a few points on how strategic reasoning in the context of extensive games can be analyzed from the perspective of modal logics of belief change.

Chapter 6

Strategic reasoning[1]

We applied the logical framework for belief change developed in Chapter 2 and 3 to the analysis of interactive reasoning (Chapter 4) and of inductive learning (Chapter 5). In this chapter we bring temporal doxastic and dynamic doxastic logics to the study of strategic reasoning in the context of extensive form games of imperfect and perfect information. We focus mainly on a more temporal approach before more briefly discussing a more dynamic approach. But in both cases the formal systems live at the edge of these two perspectives, taking features from both of them.

6.1 Introduction

In this chapter we raise the question of the logical modeling of agents' knowledge and belief in extensive games of imperfect information. The interactive dimension of these beliefs is of crucial importance in the context of strategic interaction, but we have already discussed important features of interactive reasoning. Instead we focus on another important aspect of these beliefs: they are not only past-oriented beliefs but future-oriented beliefs or expectations, namely expectations about what other agents are likely to do at later information sets. Moreover we are interested in how agents modify their beliefs and their representation of the game as they reason about it, or as the game unfolds and players take actions.

Structure of this chapter. We start by drawing a line between notions that belong to the temporal and informational structure of the game and the ones that belong to particular ways of playing it such as beliefs and strategies. We then distinguish between an approach based on the temporal structure of the game and an approach that uses epistemic plausibility models of games and captures the evolution of the game by means of dynamic operations. We also give two working examples (Section 6.2). We look at how concretely to represent as an

[1]This chapter builds on ideas from Dégremont and Zvesper [65].

ETL model the informational process of an extensive game with imperfect information and perfect recall (Section 6.3.1) and define a matching logic of programs with epistemic features to reason about the game structure (Section 6.3.2). We continue by exploring how solution algorithms for extensive games of imperfect information can be accounted for as model-changing operations representing the strategic reasoning carried out by the agents, and we extract the matching notion of rationality implicitly at work in these algorithms (Section 6.4). We then turn to doxastic languages. We first give a partial DETL perspective on an abstract notion of equilibrium using past-oriented beliefs (Section 6.5) and then turn to modeling expectations or future-oriented beliefs in imperfect information contexts, discussing assumptions and properties one could require them to fulfill and their matching syntax (Section 6.6). We then move on and put the first stone for an application of our earlier protocol-based dynamic approach to a classical case: the analysis of the role of belief change in the epistemic foundations of backward induction (Section 6.7). We conclude in Section 6.8.

A terminological remark: we use 'agents' and 'players' to mean the same thing.

6.2 Game structure and actual play

In the usual game-theoretical perspective two dimensions are important to distinguish: the description of the game and how it can be played.

The game structure encodes the protocol of the interactive situation: what players can do (the action structure of the game), what are the preferences of the players about the possible outcomes about the game (the payoff structure of the game) and what the players would know about the sequence of actions taken so far by the other players when making their decision. The game structure is a full specification of an abstract representation of an interactive, strategic decision-making problem.

But as such the game structure does not yet explain a class of observed behaviors, i.e. how real agents actually play these games, if one is concerned with a descriptive approach [54]. Nor does it determine the reasonable ways of actually playing these games, i.e. how idealized agents would play them, if one is concerned with an analytical approach [127, 124]. (We could have said "how idealized agents *should* play them", since how an idealized agent would behave is sometimes not clear and might be disputed.) Before explaining or determining as reasonable (or unreasonable) particular ways of playing a game (henceforth *plays*), let us say that a play includes a specification of the *beliefs* and *strategies* of the players. In this chapter a strategy for an agent i will be a function assigning to each of her information cells one of her available actions (intuitively the action she would take, were she to actually reach that information cell). This being said, considering the two interpretations of game theory mentioned in Section 1.5.1, beliefs might have one of two different status. Under the evolutive interpretation

they are taken to be arising from experience of past play and the main issue is to determine which profiles of strategies and beliefs are in a stable state (equilibrium analysis). Under the epistemic interpretation one is usually interested in assumptions about the players' beliefs (such as assuming that the players believe in each other's rationality) and the properties the actual play or the actual profile of strategies will satisfy given that agents' beliefs satisfy these properties. Still from a logical point of view they can be tackled in similar ways.

In general one will be interested in modeling three things from a logical perspective: the game structure, the beliefs and strategies of the agents, and how these beliefs and strategies might evolve as they receive new information from the environment (such as observing an action from another agent) or reason about the interactive situation. Analyzing strategic reasoning from a logical perspective, we encounter a wide range of notions and ideas that constitute building blocks of a study of strategic reasoning processes — i.e. how we reason about what the other agents might have done so far, about what they will do next and what they would themselves believe about the other agents' beliefs and strategies. All these notions call for specific logical developments, each of which brings a challenge of its own.

On the model-theoretic side, a few options are available and have been considered in the logical literature concerned with strategic reasoning. They can be divided in two main lines:

- Having the game structure itself or some richer temporal structure as the basis for our models (Section 6.3);
- Working with epistemic plausibility models and taking information, including moves (actual decisions), to be interpreted as model-changing operation (Section 6.8).

But we immediately stress that the division between the two approaches is not clear-cut. In the analysis of strategic reasoning they each benefit from importing features from the other. In each case we will be interested to see 'how much' strategic reasoning can be logically analyzed in these frameworks and what languages are required to do so.

It will be useful in this respect to have two concrete examples of games. They can both be represented as extensive games of imperfect information, but they are more precisely described as a *BoS* game (Example 6.1) and as a signaling game (Example 6.2). Note that in our examples numerical payoffs are just a convenient way to encode the preference orderings of the agents.

Example 6.1 (Bartók or Strauss)**.** *This example is a qualitative version of the extensive form* BoS *game. 1 and 2 have only enough money to buy one more opera ticket this season. But they have different preferences. 1 would prefer to see Bartók's* Bluebeard's Castle *and 2 would rather see Strauss'* Salomé*. However*

they prefer to attend together the opera they like less rather than going alone. 2 *was on his way to buy his ticket from the box office when his mobile ran out of batteries. Seats are being reserved very quickly, so the protocol is that* 1 *should buy her ticket online. She will get to choose first (Bartók or Strauss) and when* 2 *gets to the box office, he will also choose (Bartók or Strauss), knowing that* 1 *has reached a decision but not knowing which one.*

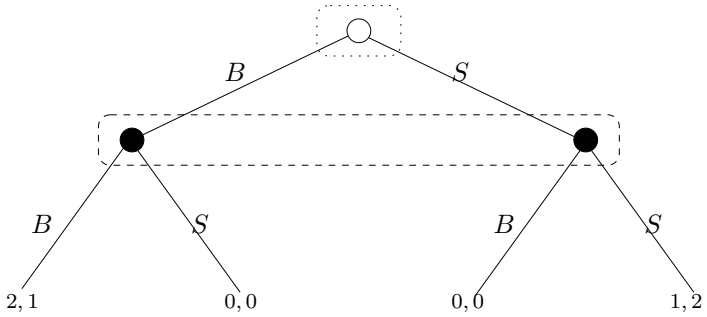

Figure 6.1: Bartók or Strauss in extensive form.

Example 6.2 (Education as Signaling)**.** *This example is a qualitative version of Spence [145]'s modeling of education (in its relation to the labor market, and more precisely to being hired and hiring) as a signaling game. (More precisely it is a qualitative version of the setting used in Brandts and Holt [52].)*

An individual W (worker) has a type L or H (representing her skills, intelligence or taste for hard work) that might be either Low or High. Only she knows it. She will get to choose whether to invest (I) in or skip (S) college education. Observing the decision W has made (I or S) but not her type, an employer E will decide whether to assign W to a challenging (C) or dull (D) job. Type L agents would prefer to skip college, but would still prefer going to college and getting a C job rather than skipping it and getting the D job. In short: $(C, S) >_L (C, I) >_L (D, S) >_L (D, I)$*. An agent of type H is firstly concerned about getting a C job, but given a type of job, she prefers to have received college eductation:* $(C, I) >_L (C, S) >_L (D, I) >_L (D, S)$*. The employer's unique concern is to match L-type workers with D(ull)-jobs and H-type workers with C(hallenging)-jobs. (See Figure 6.2.)*

Let us now see how to represent such games as models on which we will be able to interpret the doxastic temporal and dynamic doxastic logics we have met earlier as well as new interesting members of this family.

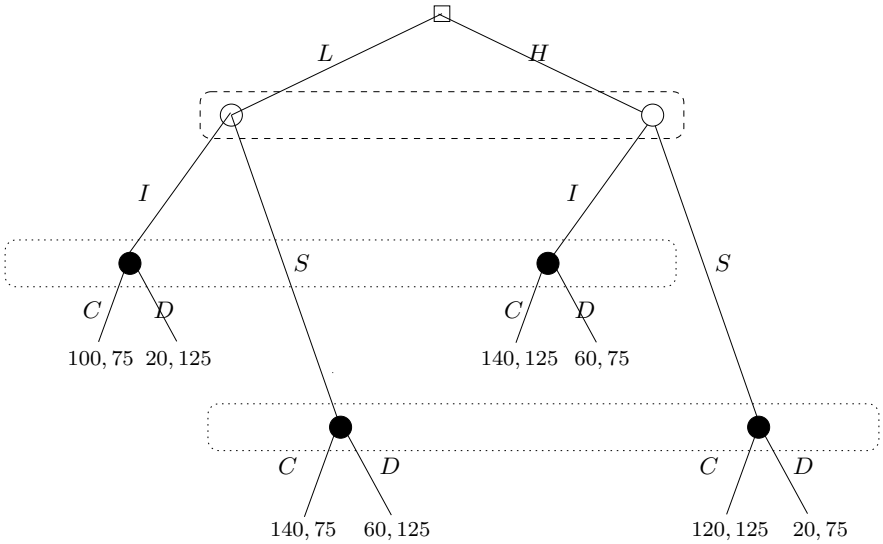

Figure 6.2: Modeling education as a signaling game.

6.3 Using the game structure as a model

It is tempting to have the game structure itself as the base of our doxastic epistemic temporal models. Indeed extensive games of perfect information are really process models [26, 46]. And extensive games of imperfect information are really process models with an uncertainty relation (van der Hoek and Pauly [105]): the process and informational structure of the extensive form game (with imperfect information) gives the temporal and epistemic part of our DETL models. By adding plausibility relations one can furthermore encode to some extent the beliefs of the agents. But both knowledge and beliefs will be past-oriented, beliefs and knowledge about what has happened so far.

We start by showing concretely how to encode a finite extensive form game into an epistemic temporal model. The details of the encoding are different from but related to the representation used in van der Hoek and Pauly [105]. We then consider different notions of belief that are relevant for players' strategic reasoning in extensive games with imperfect information, illustrating how they correspond to different belief operators in some modal logical languages.

6.3.1 Extensive games of imperfect information as epistemic temporal models

At a first glance, it might seem straightforward to transform an extensive game of imperfect information into an ETL model. But, in the general case, i.e. for an arbitrary extensive game of imperfect information, it is not possible to complete its information structure in a way that preserves certain natural consistency and

memory conditions (for such conditions read further, and for such impossibility results using similar assumptions see Quesada [135] and also Battigalli and Bonanno [23], Bonanno [45]). As a consequence there is sometimes no satisfactory way to construct an ETL model from a game. Therefore we restrict our attention to finite von Neumann games.

Definition 6.3 (von Neumann games, Battigalli and Bonanno [23])**.** *An extensive game with imperfect information is a* von Neumann game *if, whenever two nodes belong to the same information set, these two nodes have the same number of predecessors.*

To encode the structure of an extensive game with imperfect information (cf. Definition 1.23) into an epistemic temporal model we proceed as follows.

Definition 6.4 (ETL model generated from an extensive game form with imperfect information and perfect recall)**.**
Defining the set of events*. We define our set of events Σ to be $Act \times N$, i.e. pairs of the form (a, i), whose intuitive sense is that player i takes action a.*

Defining the set of histories*. Since by definition of extensive-form games the edges going from a given node to its successors are labelled with unique actions, every sequence of actions can be mapped to at most one node in the extensive form games. We can recursively define the set H of histories together with a mapping from H to T as follows.*

We let $H_0 := \{\epsilon\}$, where ϵ is the empty sequence and define $f(\epsilon) := t_0$ where t_0 is the root of the extensive-form game. Now assume that $h \in H_n$, $ha \notin H_n$ and that $f(h) = t$, $\rho(t) = i$, $t < s$ and $A(t, s) = a$; we define $H_{n+1} := H_n \cup \{ha\}$ and define $f(ha) = s$. We then define our set of histories to be $H := \bigcup_{n\in\omega} H_n$.

Defining the uncertainty relation*. We can now define the uncertainty relation $h \sim_i^0 h'$ iff $f(h) \equiv_i f(h')$. Note that we cannot stop there since $\sim^0$ is not yet an equivalence relation on H. To make it so, one way to go is to take the smallest equivalence relation containing $\sim_i^0$ still satisfying Perfect Recall (Definition 2.1), knowledge of one's turn and of one's available actions (Definition 6.5) and memory of one own choices (Definition 6.6). This corresponds to Battigalli and Bonanno [23]'s notion of* maximal information*. (Let us remark that for arbitrary extensive games of imperfect information, there might no such equivalence relation, hence the restriction to von Neumann games.)*

Definition 6.5 (Knowledge of one's turn and of one's available actions)**.** *In the $ETL^{\succeq^z}$ model of a game, i has knowledge of her turn and available actions iff for every h, h' with $h \sim_i h'$, if $\exists h''$ with $h' = h''(a, i)$ then $\exists h'''$ with $h = h'''(a, i)$.*

Definition 6.6 (Memory of one's own choices)**.** *In the $ETL^{\succeq^z}$ model of a game, i has memory of her own moves iff, if $(h(a, i) \sim_i h')$ then $\exists h''$ with $h' = h''(a, i)$.*

Now to analyze strategic reasoning it will be necessary to reason about what the agents' preferences are. The natural extension of ETL models is as follows:

Definition 6.7 (ETL model with terminal preferences)**.** *An ETL model with terminal preferences ($ETL^{\succeq^z}$) is an ETL model together with a total (preference) pre-order $\succeq_i^z$ on maximal histories.*

Given a particular game, the $ETL^{\succeq^z}$ model generated from a game is defined in the obvious way, by taking the preference relation to be a copy of the total pre-order given in the game.

Finally one might like to consider an extended preference relation on non-terminal nodes. We will explain in detail the motivation to do so. But in short, it allows to work with a local concept of rationality, local in the sense that it can be defined by a modal formula whose interpretation requires to scan only a bounded part of the game tree. Also such lifting of preferences from leaves in the direction of the root is, more or less explicitly, at work in solution algorithms such as backward induction. We get back to this issue in Section 6.4.

As an illustration, the $ETL^{\succeq^z}$ model generated from the game of example 6.1 (Figure 6.1) is displayed in Figure 6.3, letting the preference relations to be encoded by a payoff function for easier readability. We leave out the uncertainty relation between end nodes since it is irrelevant for strategic reasoning.

Now that we have our models let us turn to the matching languages and see 'how much' strategic reasoning then can be expressed.

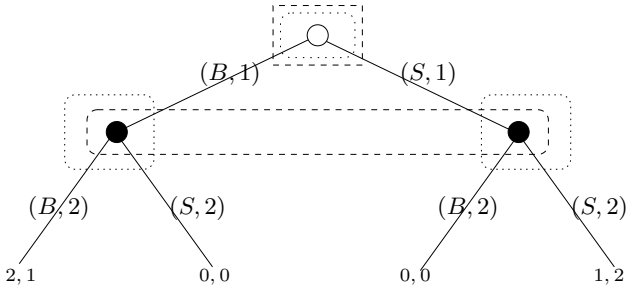

Figure 6.3: Bartók or Strauss in extensive form as $ETL^{\succeq^z}$ model.

6.3.2 Reasoning about games with a logic of programs

We introduce a first language $\mathcal{L}_{GDL}$ that can be used to check whether a given model encodes properly the situations one had in mind or to define classes of frames corresponding to properties one would expect $ETL^{\succeq^z}$ models of games to have. This language does not include anything about beliefs. It is maybe best described as a boolean epistemic language since it does not contain iteration, but we stick with the process model analogy. Its syntax follows.

Definition 6.8 (A PDL style language to reason about games)**.** *Our language has a recursively defined set of programs:*

$$\alpha ::= (a, i) \mid a \mid i \mid{\rightarrow}\mid \succeq_i \mid \succ_i \mid \sim_i \mid \alpha \cup \alpha \mid \alpha;\alpha \mid \alpha \cap \alpha \mid \alpha^{-1}$$

To each program corresponds a modality $\langle\alpha\rangle$ *in the language* $\mathcal{L}_{GDL}$*:*

$$\varphi ::= p \mid \neg\varphi \mid \varphi \vee \varphi \mid \langle\alpha\rangle\varphi \mid K_i\varphi \mid C_G\varphi$$

where i *ranges over* N*,* G *over* $\wp(N) \setminus \{\emptyset\}$*, and* p *over proposition letters* PROP*.*

The semantics of $\mathcal{L}_{GDL}$ is interpreted over nodes h in $ETL^{\succeq_z}$ generated from extensive form games of imperfect information. We start by giving the interpretation of programs:

Definition 6.9 (Interpretation of programs)**.**

$$\begin{array}{lcl}
R_{\sim_i} & = & \sim_i \\
R_{\succeq_i} & = & \succeq_i \\
R_{\succ_i} & = & \{(h,h') \mid (h,h') \in \succeq_i \text{ and } (h',h) \notin \succeq_i\} \\
R_{(a,i)} & = & \{(h,h') \in H \times H \mid h' = h(a,i)\} \\
R_a & = & \{(h,h') \in H \times H \mid h' = h(a,i) \text{ for some } i\ \} \\
R_i & = & \{(h,h') \in H \times H \mid h' = h(a,i) \text{ for some } a\ \} \\
R_{\rightarrow} & = & \{(h,h') \in H \times H \mid (h,h') \in R_a \text{ for some } a\ \} \\
R_{\alpha\cup\beta} & = & R_\alpha \cup R_\beta \\
R_{\alpha;\beta} & = & \{(h,h') \in H \times H \mid \exists h'' \text{ with } (h,h'') \in R_\alpha \text{ and } (h'',h') \in R_\beta \\
R_{\alpha\cap\beta} & = & R_\alpha \cap R_\beta \\
R_{\alpha^{-1}} & = & \{(h,h') \in H \times H \mid (h',h) \in R_\alpha\ \}
\end{array}$$

Now that we have given the interpretation of the programs, the interpretation of our language is given by the two following clauses:

Definition 6.10 (Truth definition)**.** *Let* $\mathcal{K}_i[h] = \{h' \mid h \sim_i h'\}$*.*

$$\begin{array}{lll}
\mathcal{H}, h \Vdash \langle\alpha\rangle\varphi & \text{iff} & \text{for some } h' \text{ with } hR_\alpha h' \text{ we have } \mathcal{H}, h' \Vdash \varphi \\
\mathcal{H}, h \Vdash K_i\varphi & \text{iff} & \text{for every } h' \text{ with } h' \in K_i[h] \text{ we have } \mathcal{H}, h' \Vdash \varphi
\end{array}$$

On the level of pointed models, one can use this language to check the correctness of the modeling of a given scenario. As an illustration, our first scenario required, at the beginning of the game, common knowledge between 1 and 2 that whatever opera 1 buys a ticket for, when 2 will take his decision he won't know what decision 1 has reached. Using the preceding language one can verify that the model is indeed correct. To do so we simply check that at the empty sequence in the $ETL^{\succ}$ model of Figure 6.3, we have $\mathcal{H}, \epsilon \Vdash C_{\{1,2\}}[B \cup S]\neg(K_2\langle B^{-1}\rangle\top \vee K_2\langle S^{-1}\rangle\top)$.

On the level of frames one can determine syntactic correspondents to certain assumptions about the game and the agents giving us insight into the logic of informational flow in games and its specificities. As an example let us look at the definability of the earlier properties relating to the game-theoretical notion of perfect recall. The analysis in this section is closely related to the one in Bonanno [44] which uses a slightly different language and slightly different model-theoretic primitives. We start with 'memory of one's own choices' (Definition 6.6). But first, let us recall how the determinacy of actions can be characterized:

Fact 6.11 (Defining determinacy, see e.g. [39])**.**
The formula $\langle a\rangle p \to [a]p$ *characterizes the class of action-deterministic frames.*

Now on the class of action-deterministic frames, memory of one's own choices can be characterized as follows:

Fact 6.12 (Defining memory of one's own choices)**.**
On the class of action-deterministic ETL frames, $\langle (a,i)^{-1}\rangle \top \to K_i\langle (a,i)^{-1}\rangle \top$ *characterizes memory of one's own choices.*

Now if we assume perfect recall (in the sense of Definition 2.1) we can use a stronger formula, which does not require the converse construct.

Fact 6.13 (Defining memory of one's own choices on frames with perfect recall)**.**
On the class of action-deterministic ETL frames with perfect recall,

$$\langle (a,i)\rangle\langle \sim_i\rangle p \to \langle \sim_i\rangle\langle (a,i)\rangle p$$

characterizes memory of one's own choices.

The same phenomenon can observed for the definability of perfect recall itself with respect to the assumption of synchronicity. But, in general, extensive games with imperfect information and (game-theoretical) perfect recall need not satisfy synchronicity. So the stronger axiom ($\langle \to\rangle\langle \sim_i\rangle p \to \langle \sim_i\rangle\langle \to\rangle p$) would characterize a smaller class of games.

On the technical side, these schemes of axioms are usual suspects in the epistemic temporal literature as they can be used to force grid-like structures. This comes as bad news since the corresponding modal logics are rapidly undecidable and even non-finitely axiomatizable. In fact it follows from a theorem by Halpern and Vardi [94] that adding iteration (and test) to $\mathcal{L}_{GDL}$ would immediately exclude the possibility of axiomatizing its validities over the class of ETL frames with perfect recall. For the big picture about such complexity results for epistemic temporal languages, the reader can check [34].

6.4 Solution concepts changing the models

But this is not the end of what a non-doxastic language can do. It is indeed possible to model a bit of strategic reasoning in such structures. The idea is that the model itself encodes the representation agents have of the interactive situation. If we're ready to accept the idea that pre-play reasoning is transparent in the sense that every agent will reason in the same way and expect everyone to expect everyone else to reason in the same way, then one can focus on this common reasoning directly on the level of ETL models.

To do so, van Benthem [31] develops the analogy between a solution concept and a model-change operation such as public announcement. Indeed eliminating part of a game as incompatible with a given solution concept is modifying one's representation of an interactive situation, and can thus really be seen as eliminating part of a model that is incompatible with some rationality concept. The process of applying a solution algorithm such as iterated strict dominance can really be thought of as the process of assuming that agents will behave in certain (rational) ways and thus as eliminating the possibilities that the agents will behave in certain other (irrational) ways. Moreover applying iterated strict dominance to the game can only be a correct way of simulating agents' reasoning on the condition that this process is transparent and commonly 'known' to be performed by all agents.

But now given an epistemic representation of a strategic game in its natural $\mathbf{S5}^N$ product form (see [31] for details), it is possible to define a formula φ such that the public announcement of φ in the epistemic model of the strategic game mimics exactly the effect of applying strict dominance once. Intuitively φ is the notion of rationality underlying the concept of iterated strict dominance: a player won't select an action that she knows to be strictly dominated. And a public announcement of rationality eliminates all parts of the (epistemic representation of the) game that correspond to a play in which at least one agent is doing something irrational [31].

That the resulting model is still a(n epistemic representation of a) game is guaranteed by the fact that '(the chosen notion of) irrationality is introspective', i.e. if at some state i is irrational, then i knows she is irrational. In this way if a state is eliminated, it is because some agent j would be irrational in the corresponding play of the game, but then by introspection, the whole information cell for j (which corresponds to an action for j) will be eliminated. Finally the idea of a public announcement of rationality might sound counterintuitive, but in the DEL way of thinking the event that all agents remove all states of the model that do not satisfy φ ('rationality') in such a way that this process is transparent and commonly 'known' to all agents, is exactly what a public announcement of φ ('rationality') is doing.

But the crucial point is that, as indicated by van Benthem [31], to different solution concepts will really correspond different notions of rationality. We have

explained this notion of correspondence between a notion of rationality and a solution concept on an intuitive level so far. For the case of extensive games, there is more than one possible way to define the correspondence and we are not claiming to give the 'right' one. Instead we give a relatively general definition, that needs to parametrized. The parameters will be discussed right after the definition.

Definition 6.14 (Rationality concept corresponding to an iterative solution concept)**.** *Given a class of extensive games (of imperfect information) $\mathcal{C}$: a notion of rationality φ corresponds to an iterative solution concept $\mathcal{SOL}$ on $\mathcal{C}$ iff for every game $\mathcal{G} \in \mathcal{C}$ the following sets have the same extension:*

1. *The set of profiles of pure strategies in $\mathcal{SOL}(\mathcal{G})$.*
2. *The set of profiles of pure strategies that can be defined within the fixed point of the ETL$^{\succ}$ model of a game generated by $\mathcal{G}$ under publicly announcing φ and then closing under preferences.*

The first parameter concerns the class of games: usually games of perfect recall, sometimes of perfect information, sometimes without ties in payoff, to cite a few. The second parameter is the notion of preference-closure we are using to lift preferences from terminal nodes towards the root. Another option — which is a syntactic way to look at preference closure — is to encode the implicit notions of preference closure within the notions of rationality. But doing so has two drawbacks. The first is that it tends to obfuscate the underlying local notion of a 'rational decision' by merging it in a single formula together with that of a notion of temporal coherence of preferences. The second is that it tends to push definability out of the scope of decidable logics.

This definition and these comments are fairly abstract. It might be useful to give concrete examples. We start with the case for which things are the clearest. We will give a paradigmatic local notion of rationality putting together two slogans:

Slogan 6.15 (Locality; Blackburn et al. [39])**.** *Modal languages provide an internal, local perspective on relational structures.*

Slogan 6.16 (Rationality; [126])**.** *The action chosen by a decision-maker is at least as good, according to her preferences, as every other available action.*

In fact it is easier to define a notion of 'irrationality'.

Definition 6.17 (Local irrationality in perfect information situations)**.** *We say that an agent i has just been irrational in a perfect information situation if*

$$\mathcal{H}, h \Vdash \langle (i^{-1}; i) \cap \prec_i \rangle \top$$

This formula says something simple, namely that the agent just did something such that she could have done something else that would have led her to some state that she would now prefer to the current state. Our notion of rationality of i is then just its negation: $[(i^{-1}; i)\cap \prec_i]\bot$.

Interestingly this very simple notion of rationality together with a natural notion of preference lifting corresponds to backward induction on the class of extensive games of perfect information without ties in payoff.

Proposition 6.18. *On the class of extensive games of perfect information without ties in payoff, using a natural notion of preference lifting,* $\bigwedge_{i\in N}[(i^{-1}; i)\cap \prec_i]\bot$ *corresponds to backward induction.*

Sketch. The idea is to show that announcing $\bigwedge_{i\in N}[(i^{-1}; i)\cap \prec_i]\bot$ and lifting preferences mimics exactly the backward induction algorithm (BI) on games of perfect information without ties in payoff. With ties in payoff the two differ, since the backward induction algorithm will branch, isolating only the subgame perfect equilibria, while the algorithm corresponding to announcing $\bigwedge_{i\in N}[(i^{-1}; i)\cap \prec_i]\bot$ will be weaker and accept all the possible profiles of strategies that be defined on the union of the set of subgame perfect equilibria.

Let us now see that the two algorithms match. Take the first stage, there are only preference relations on the leaves. So $\langle(i^{-1}; i)\cap \prec_i\rangle\top$ (irrationality) can only hold at one of the leaves. Now BI will start by checking the subgames of size 1 (size is counted from the leaves). If backward induction eliminates a state, then it is because it would correspond to a suboptimal decision, but then it is easy to see that $\langle(i^{-1}; i)\cap \prec_i\rangle\top$ would hold at the leaf corresponding to that decision, so, after the announcement, this leaf (and the corresponding action-edge) will be removed.

Now for stage 2 preferences will only be lifted one level, and what matters is to check that the notion of preference lifting we are using matches the one at work in the backward induction algorithm. QED

In the preceding sketch we left the notion of 'preference lifting' undefined. We think of this notion as corresponding to an idea of 'temporal coherence of preferences'. Discussing an idea that has repercussions in decision theory and philosophy of action would require (at least) another chapter. We rather leave it unanalyzed and focus instead on different (local) notions of rationality.

Still, another strategy is to encode both the temporal coherence of preferences and the rationality of decision-making in a 'global' notion of rationality.

Definition 6.19 (Momentaneous rationality; van Benthem [31])**.** *Momentaneous rationality (MR) says that at every stage of a branch in the current model, the player whose turn it is has not selected a move whose available continuations all end worse for her than all those after some other possible move.*

This is the road originally followed in van Benthem [31] who proved a similar correspondence result:

Proposition 6.20 (van Benthem [31])**.** *On the class of extensive games of perfect information without ties in payoff, MR (without preference lifting) corresponds to backward induction.*

We save the reader from a very intricate characterization of MR in a modal language using both features of hybrid logics with binders and iteration. It is indeed so demanding that it is certainly more naturally characterized as a fragment of first-order logic with transitive closure.

Let us move to the case of extensive games of imperfect information. This a place where the distinction between the epistemic and the evolutive interpretation of games will be useful. Indeed equilibria of games of imperfect information are generally expressed in terms of mixed strategies and matching probabilistic beliefs. It is not clear to us what would be the natural counterpart of such strategies in a qualitative setting. But under the epistemic interpretation solution algorithms are at least as important as equilibrium notions, and it is in our opinion somewhat easier to think of qualitative solution algorithms than qualitative equilibrium notions for games of imperfect information. In general, a qualitative theory of strategic decision-making cannot be as fine-grained as a probabilistic one, but as a consequence eliciting the information basis required in the probabilistic setting assumes much more from the agents. Moreover there are still a few natural qualitative solution algorithms whose corresponding rationality concepts can be investigated.

One of the weakest algorithms one can think of is strict dominance applied directly, backwards on the tree. In words it eliminates actions that an agent would know to be dominated at her information set. Consider the following notion of rationality.

$$\neg \bigvee_{a \in A_\Sigma(i)} \langle (a,i)^{-1} \rangle \bigvee_{b \in A_\Sigma(i)} K_i \langle ((b,i); \succ) \cap (a,i) \rangle \top \qquad (R_{SD}(i))$$

First note that, as one would expect, knowledge now appears in the formula. Moreover explicit reference to actions appear too, to express the idea of domination of an action. Finally note that the size of formula will be quadratic in the number of actions.

In what follows *extensive-form* strict dominance is a qualitative counterpart to Pearce [133]'s procedural definition characterization of extensive-form rationalizability (EFR) referred to, in Battigalli [22], as backward iterated dominance and for which Battigalli [22] provides an alternative belief-based formulation that he proves to be equivalent. Intuitively, *extensive-form* strict dominance is simulated on the level of the game model by reading, at each stage, the current support off the current sub-model and similarly for the currently surviving strategies.

Proposition 6.21. *On the class of extensive games of imperfect information without ties in payoff, using a natural notion of preference lifting,* $\bigwedge_{i\in N}(R_{SD}(i))$ *corresponds to extensive-form strict dominance.*

Sketch. The idea of the proof is the same as for Proposition 6.18. QED

Strict dominance will eliminate actions that an agent knows to be dominated at her information set. But this is as far as it goes. A stronger concept corresponding to a cautious type of decision-making is to use the *MinMax* approach. The general idea is simple: choose actions that minimize your maximum possible loss. But in a strategic context, applying it properly cannot be done by brute force. Indeed it might conflict with either domination arguments or forward induction arguments.

It would take us out of the scope of this chapter to discuss the design of a satisfactory algorithm using *MinMax* for extensive games of imperfect information in a qualitative setting. But the underlying notion of rationality of the pure *MinMax* algorithm is the following:

$$\neg \bigvee_{a\in A_\Sigma(i)} \langle (a,i)^{-1}\rangle \bigvee_{b\in A_\Sigma(i)} (\langle (b,i)\rangle \top \wedge \langle \sim_i ; (a,i)\rangle [(\succeq_i ; (b,i)^{-1}) \cap ((a,i)^{-1} ; \sim_i)]\bot \tag{6.1}$$

Similar questions arise for an approach based on weak dominance. But again the underlying notion of rationality of the pure weak dominance algorithm is the following:

$$\neg \bigvee_{a\in A_\Sigma(i)} \langle (a,i)^{-1}\rangle \bigvee_{b\in A_\Sigma(i)} ((K_i\langle ((b,i);\succeq)\cap (a,i)\rangle \top) \wedge (\langle \sim_i\rangle \langle ((b,i);\succ)\cap (a,i)\rangle)) \tag{6.2}$$

A syntactic counterpart to the fact that weak dominance is stronger than strict dominance, in the sense it will eliminate more strategies, is that

$$K_i\langle ((b,i);\succ)\cap (a,i)\rangle \top \rightarrow ((K_i\langle ((b,i);\succeq)\cap (a,i)\rangle \top) \wedge (\langle \sim_i\rangle \langle ((b,i);\succ)\cap (a,i)\rangle))$$

is valid, under the simple assumption that the uncertainty relation is serial, i.e. that knowledge is always consistent.

6.5 Past-oriented beliefs in equilibrium

So far we have discussed knowledge-based approaches to strategic reasoning. Other aspects of strategic reasoning involve different notions of *beliefs*, which are properties of plays or of an equilibrium, but not given by the structure of the game. For example we might precisely be interested in checking if certain profiles of strategies and beliefs are in equilibrium or if they satisfy certain optimality properties. To do so the first step is to encode the beliefs of the agents. How

much can be done without drastically transforming the given epistemic temporal structure?

As the reader might expect, we will first consider moving to doxastic epistemic temporal models. Doing so allows us, by extending the epistemic temporal model generated from the game with plausibility relations, to encode past beliefs of the players, i.e. beliefs about previous actions taken so far. Let us show an application of this simple DETL perspective by giving a starting point to a logical characterization of a simple notion of equilibrium. To start with, compare the following three plays of the game in Example 6.1 (Figures 6.4 to 6.6).

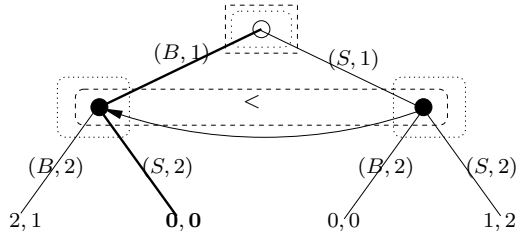

Figure 6.4: A first play of the Bartók or Strauss game.

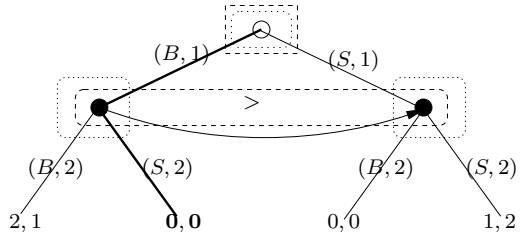

Figure 6.5: A second play of the Bartók or Strauss game.

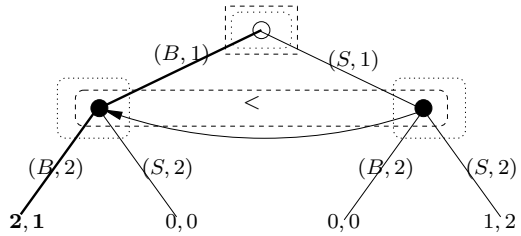

Figure 6.6: A third play of the Bartók or Strauss game.

Of these three plays only the third (Figure 6.6) is an equilibrium. The first play is not an equilibrium because player 2's decision at his information set is not

a best response to his beliefs. The second play is not an equilibrium because his beliefs about 1's strategy are incorrect. If the game is to be played repeatedly like this, his beliefs would change. In the third example his beliefs are in equilibrium with 1's strategy and his decision is a best response. To complete the equilibrium analysis we would need to look at things from the perspective of 1 as well, but let us get come to it. For now we can already extract two important ideas for a logical characterization. A play is an equilibrium if

- agents' beliefs are correct and
- if their decisions are best responses to their beliefs.

But for past-beliefs these are notions that can be characterized by natural DETL formulas, with as usual $P\varphi \leftrightarrow \langle(\rightarrow^{-1})^*\rangle\varphi$:

Fact 6.22 (Correctness of past-beliefs)**.** *Player i's past-beliefs are correct throughout the play h of the game if*

$$\mathcal{H}, h \Vdash P \bigwedge_{(a,j)\in\Sigma} B_i\langle(a,j)^{-1}\rangle\top \rightarrow \langle(a,j)^{-1}\rangle\top$$

Fact 6.23 (Best response to past-beliefs)**.** *Player i's actions are optimal with respect to her past beliefs throughout the play h of the game if*

$$\mathcal{H}, h \Vdash P \bigwedge_{a\in A_\Sigma(i)} [(a,i)^{-1}]\neg B_i \bigvee_{b\in A_\Sigma(i)} \langle(b;\succ)\cap(a,i)\rangle\top$$

This gives us a starting point for a logical characterization of equilibrium notions. To complete the analysis we now need to look at future beliefs. Indeed our decisions need not only to be best responses to our beliefs about the past but also to our expectations about what others will do after us.

6.6 Future we can believe in

In the same way that our beliefs about the past can evolve as we observe events, so can our beliefs about the future, about what will happen next. And these beliefs about the future (expectations) might need to be revised when they are defeated by new information.

The information partition encodes what players know about previous actions, about the play so far. A plausibility ordering selects the most plausible elements of a cell of information, indicating what are the agents' beliefs (and conditional beliefs) about what has happened so far. Without changing radically the underlying model-theoretical representation of the interactive situation, one can consider richer structures, building on DETL models, on which one can interpret languages that allow for future beliefs or *expectations*.

Before we give the details, we need a bit of notation. Given an *ETL* model, let Z be the set of maximal histories and $\mathbf{H}(h) = \{\ h' \in Z \mid h \sqsubseteq h'\ \}$ be the set of maximal histories still open at h. (As before $h \sqsubseteq h'$ means that h is a prefix of h'). Finally given a set H of histories, we write $\mathbf{H}(H) := \bigcup_{h\in H} \mathbf{H}(h)$. We can now define the notion of expectation function.

Definition 6.24 (Expectation function)**.** *The agents' beliefs about the future history or expectations will be defined by an expectation function* $EX_i : H \to \wp(Z)$ *such that* $EX_i(h) \subseteq \mathbf{H}(\mathcal{K}_i[h])$.

Intuitively an expectation function selects the histories that i considers to be the most likely continuations. An *ETL* model with expectations is then an *ETL* model with a profile of expectation functions.

As for assumptions about expectation functions, it seems natural to require that agents are introspective with respect to their expectations. Formally we require:

Definition 6.25 (Introspection of expectations)**.** *In an ETL model with expectations, an agent* i *satisfies* introspection of expectations *iff* $h \sim_i h'$ *implies* $EX_i(h) = EX_i(h')$.

In the same way as for *ETL* models, a *DETL* model with expectations is simply a *DETL* model with a profile of expectation functions. In this context, a natural consistency property one might require is that expectations are consistent with (past oriented) beliefs.

Definition 6.26 (Consistency of expectations with beliefs)**.** *In DETL model with expectations, an agent* i *satisfies* consistency of expectations with beliefs *iff* $EX_i(h) \subseteq \mathbf{H}(\mathcal{B}_i[h])$ *where as usual* $\mathcal{B}_i[h] = \min_{\leq_i}(\mathcal{K}_i[h])$.

The idea is that an agent should expect histories to happen that are compatible with what she believes the past history to have been so far.

Concerning the logic itself we add a branching-time type operator (in the sense that it quantifies on future continuations) $[\Upsilon_i]$ for each agent i. $[\Upsilon_i]\varphi$ means that i expects that φ or more precisely that in all the future continuations that i considers the most likely, φ is holds. It has the following branching time semantics (cf. Section 3.3.2):

$$\begin{array}{rl} \mathcal{H}, \epsilon, h \Vdash [\Upsilon_i]\varphi & \text{iff } \forall \epsilon' \in EX_i(h)\ \ \forall h' \text{ such that } h' \sqsubseteq \epsilon' \text{ and } h' \in \mathcal{K}_i[h] \\ & \text{we have } \mathcal{H}, \epsilon', h' \Vdash \varphi \end{array}$$

The two constraints on the relation between expectations and both knowledge and (past-oriented) beliefs that we have been considering have their natural syntactic correspondents:

$$[\Upsilon_i]\varphi \ \to\ K_i[\Upsilon_i]\varphi \tag{6.3}$$

In words: if i expects that φ, then i knows that she expects that φ. And

$$B_i \forall \varphi \ \rightarrow [\Upsilon_i]\varphi \tag{6.4}$$

If i believes that in all possible continuations φ holds, then she expects that φ.

Note that we cannot require the converse direction, since the whole idea behind expectations is that they select a subset of all the possible future courses of action that an agent believes to be still open.

Further constraints include consistency properties in the way expectations are revised. Such assumptions have been considered in the context of another approach to expectations in a sequence of papers by Bonanno [41, 42]. We first sketch out the idea behind this framework and then try to identify what these constraints would boil down to in the approach we have presented.

6.6.1 Revising expectations

Modeling predictions. Bonanno [41, 42]'s approach is based on *prediction frames* which are closely related to extensive games of perfect information and we will think of them as extending tree-like *ETL*-frames of perfect information, i.e. *ETL*-frames in which for every h and agent i, $|\mathcal{K}_i[h]| = 1$, with a prediction relation for each agent. Recall that $R_{\rightarrow}$ is really the successor relation in the *ETL* model. A prediction relation is then a sub-relation R_p of $(R_{\rightarrow})^+$, the transitive closure of $R_{\rightarrow}$. As usual $R[h]$ is the image of h under R, i.e. $R[h] = \{h' \in H \mid (h, h') \in R\}$.

Bonanno [42] proposes a principle of *minimal revision* that implements a conservativity rule requiring that if some previous predictions are still compatible with the actual course of things, then the agent should stick to these predictions, i.e. she should predict *all* the continuations that she was previously predicting that have not been defeated yet (MR1) and *only those* (MR2). It naturally splits in two parts. (MR1) is thus a principle of *non-contraction*:

$$\text{if } h_1 R_p h_2 \text{ then } R_p[h_1] \cap (R_{\rightarrow})^+[h_2] \subseteq R_p[h_2] \tag{MR1}$$

while (MR2) is thus a principle of *non-expansion*:

$$\text{if } h_1 R_p h_2 \text{ and } R_p[h_1] \cap (R_{\rightarrow})^+[h_2] \neq \emptyset \text{ then } R_p[h_2] \subseteq R_p[h_1] \cap (R_{\rightarrow})^+[h_2] \tag{MR2}$$

What would be a corresponding property in the expectation framework? In line with our previous approach, a natural way of generalizing the preceding revision properties to the general *ETL* case with expectations would be as follows: if some previous expectations are still compatible with our current *beliefs*, then the agent should stick to these expectations, i.e. she should predict *all* the continuations that she was previously predicting that are still compatible with her current *beliefs* (MER1) and *only those* (MER2).

By choosing beliefs over knowledge in the previous definitions we guarantee that the compatibility of expectations with beliefs (Definition 6.26) has priority over the conservativity of expectation revision. Formally we have the two following principles:

$$\text{if } (h_1, h_2) \in (\sim_i \circ (R_{\rightarrow})^+) \text{ then } EX_i(h_1) \cap \mathbf{H}(B_i[h_2]) \subseteq EX_i(h_2) \qquad \text{(MER1)}$$

while (MER2) is as expected the corresponding notion of *non-expansion*:

$$\begin{aligned} \text{if } (h_1, h_2) \in (\sim_i \circ (R_{\rightarrow})^+) \text{ and } EX_i(h_1) \cap \mathbf{H}(B_i[h_2]) \neq \emptyset \\ \text{then } EX_i(h_2) \subseteq EX_i(h_1) \cap \mathbf{H}(B_i[h_2]) \end{aligned} \qquad \text{(MER2)}$$

As for their syntactic correspondents, we expect similar properties as the one found in [42] and leave the question open.

Such a logic of expectations can be put to work to encode beliefs of players about what other agents are going to do. Using a conditional notion of expectation we think that it is possible to encode beliefs about strategies and also a corresponding notion of best response, and to complete the previous logical analysis. But we leave these questions open for now. Before we turn to the other possible approach, based on epistemic plausibility models and dynamic epistemic and doxastic logics, we make a remark about another temporal approach.

6.6.2 Enriching the temporal structure

Instead of working with a branching-time language with expectation operators, it is possible to work with richer process models reducing uncertainty about the future to uncertainty about the past. This line of work is very much in line with Harsanyi [98] and with the epistemic game theory literature in the sense that it steps away from the given structure of the game to work with models of games in which players' beliefs and strategies are taken as objects of uncertainty in the same way that external parameters are in Harsanyi [98]'s analysis of games of incomplete information. Among the possible logical models, epistemic plausibility models (and related models) are frequently considered, but richer process models occupy a natural position on the scale of the possible trade-offs between preservation of a natural temporal structure of reasonable size and complexity of the language.

In Figure 6.7, we illustrate such models with the game of Example 6.2 in which only some possible types of the players are considered. In particular not all strategies of Employer are considered.

Now beliefs about strategies can be encoded as past-beliefs. Here is an example of a simple piece of reasoning. At the darker node, call it w, the Worker knows her Low-type. It also the case that Worker believes it is more likely that Employer

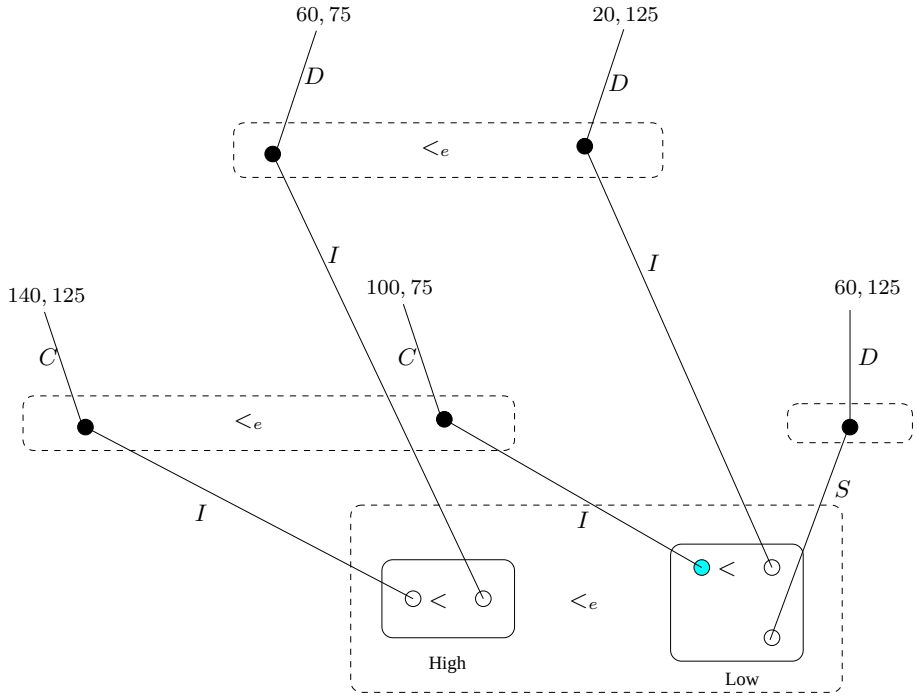

Figure 6.7: Modeling education signaling game in a richer $ETL^{\succ^z}_{-}$ model.

will give a C(hallenging) job to an agent signaling college education (I). This is expressed by the following formula:

$$w \Vdash B_w^{\langle I \rangle \top} \langle I \rangle (\langle C \rangle \top \wedge [D] \bot)$$

Worker also knows that Employer will give a D(ull) job to an agent who has skipped college (S).

$$w \Vdash K_w [S] (\langle D \rangle \top \wedge [C] \bot)$$

So Worker (correctly) believes that she is better off taking college education:

$$w \Vdash B_w^{\langle S \rangle \top} [S] G (G \bot \rightarrow \downarrow x. H (H \bot \rightarrow B_w^{\langle I \rangle \top} G (G \bot \rightarrow \langle \succ_w \rangle x)))$$

The trade-off is roughly speaking between richer process models but being able to express future-oriented beliefs (expectations) in a more standard DETL language or to extend the language with modalities quantifying over branches.

We conclude here our discussion of temporal approaches to strategic reasoning and turn to the more frequently considered, dynamic approach based on epistemic plausibility models.

6.7 A dynamic approach to strategic reasoning

So far we have considered temporal analyses of strategic reasoning. Another approach more in line with the dynamic approach works with epistemic (plausibility) models of games. They are what the reader might expect: epistemic plausibility models whose states are labeled either by complete plays or by strategy profiles. In this perspective the 'dynamic' aspects of extensive games are modeled as model change operations. We think it is interesting to present these models at work on a concrete case: the epistemic foundations of backward induction and the role of belief change in its analysis. As such the observations we will make have been made before, using different logical models. But it is interesting to see how the dynamic doxastic framework with protocols can be applied to this test case.

6.7.1 Is backward induction logical?

A sequence of papers including Stalnaker [147, 146], Halpern [92], Board [40] and Baltag et al. [21] uses doxastic, often plausibility-based models to analyze strategic reasoning in extensive games of perfect information in general, and to discuss the epistemic foundations of backward induction in particular. We think it will be useful to see how indeed an important part of the strategic reasoning at stake in the discussion of the epistemic foundations of backward induction involves the question of belief change and can be approached from a logical perspective, using epistemic plausibility models and a 'dynamic' (logic) style of thinking.

There are at least two ways to think about the problem of backward induction and its epistemic foundations. The first problem was raised by Reny [139] in a more general way. The way of putting is as follows:

Problem 6.27. *Is it possible for an agent to be better off by deviating from the backward induction solution?*

The problem is serious since if indeed a theory of games is a theory of 'rational behavior' then a solution concept, which defines what it means to be rational in a given class of games, should be

> immune to defections from it. That is, it must never be to one's advantage to behave in a manner that the theory deems irrational (Reny [139]).

But, to check if a solution concept (such as backward induction) is immune to deviations from it, we need to know what the theory recommends if someone *does* deviate and what would be the outcome of the game after the deviation. But, at least in an epistemic analysis, decisions of agents off the equilibrium-path crucially depend on their beliefs about what is going to happen next. So the theory should say something about what agents should believe after something

unexpected has happened. Indeed assume that common belief of rationality (and rationality) implies that the agents will play according to the backward induction solution. Once someone deviates should not the other agents stop believing that rationality is common belief?

This actually leads us to the second way of putting the problem (for the two player case, without ties in payoffs):

Problem 6.28. *Is it possible that there is common belief of rationality between two (rational) agents, and that they don't end up in the backward induction outcome?*

What these papers have shown (among other things) using a logical approach is that the answer depends on agents' belief revision policy. As a two-players example, if agent 1 is going to keep believing that it is common belief of rationality between them, then whatever 2 does, 2 is better off not deviating from what the backward induction solution dictates. On the other hand, if one agent (say 1) is very keen to stop believing in the other players' rationality (or maybe in the other agent's belief in her rationality), then once 2 deviates, 1 best response to her belief might be *not* to respect the backward induction solution. But then 2 might be justified in deviating in the first place. The conclusion is that for some belief revision policies, as encoded by some plausibility orderings, the out-of-equilibrium beliefs of the agents might not support anymore the backward induction play and thus, for such a belief revision policy backward induction might not be supported *at all*, even under common belief of rationality.

6.7.2 Exploring the logical dynamics of backward induction

This was a fairly abstract way of telling the story so let us give a concrete game-theoretical scenario that illustrates this discussion and explains how a logical analysis is usually brought to it. The example and the analysis follow the lines of Board [40, sec.5].

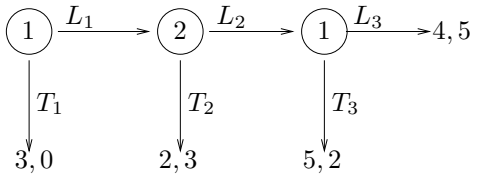

Figure 6.8: A centipede game.

This game is a 3-legs centipede game. The idea behind is that alternatively each player will get the opportunity to stop the game by taking (T) the money

on the table or to leave the money (L) on the table, in which case both agents will have two extra dollars to share. The important paramater is that you get a nicer share if you're the one taking the money (and stopping the game) and that leaving the money is only interesting for you if the other is *not* going to take the money at the next stage of the game. The left payoff is thus the payoff for player 1 and right payoff the payoff for player 2.

We will first indicate how to represent possible scenarios based on this game using epistemic plausibility models, plausibility event models and protocols. The idea is that each move in the game is information possibly triggering some plausibility change. Now the particular type of trigger depends on the scenario (i.e. of the play). Roughly speaking the representation depends not only on the game but on the personality of the players. Formally speaking, it is represented as a plausibility event model that can change the plausibility ordering of the agents.

Now the protocol maps states to sequences of pointed event models that correspond to a possible evolution of the informational process: it is constrained by both the structure of the underlying game and the particular play of the game. Finally the epistemic plausibility models encode the beliefs and knowledge of the agents about each other's strategies and beliefs, and there exists a surjective function from the domain of the model to the set of possible strategy profiles of the game. So let us look at how to model the game in Figure 6.8.

Initial epistemic-plausibility model. To lighten the model, our initial states correspond only to a subset of the possible strategies, since it is sufficient to represent this part of the model to present the main idea: $\{T, LTL, LTT, LLT, LLL\}$. We take the epistemic partition of 1 to be $\{T\}, \{LTL, LTT, LLT, LLL\}$ and LLT to be the state that 1 considers most likely. The epistemic relation of 2 is the total relation and T is the state she considers most likely.

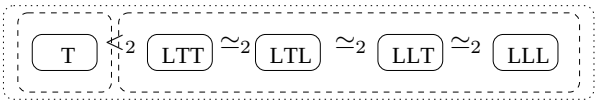

Figure 6.9: The initial epistemic plausibility models.

Modeling actions as event models. The actions T_1, T_2, T_3 and L_3 are the easiest to represent. Since they end the game immediately, we will define them as event models with a singleton domain, a total epistemic uncertainty relation and a total plausibility ordering. Since agents' beliefs at the last information set are (in such a perfect information setting) irrelevant we can model L_2 in the same way. For L_1 we will consider event models with two elements, depending on whether 1 intends to leave the money or to take it at the final information state. Let us call them L_T and L_L. Their respective plausibility ranking will be depend on the scenario we consider.

Protocol. Finally the protocol is as follows: $P(T) = \{T_1\}$, $P(LT)$ is the prefix closure of $\{L_L; T_2, L_T; T_2\}$, $P(LLT)$ is the prefix closure of $\{L_T; L_2; T_3\}$ and $P(LLL)$ is the prefix closure of $\{L_L; L_2; L_3\}$. Let us consider our two scenarios.

Preferences. The preference relation is the obvious total pre-order on the set of terminal sequences given by the protocol.

Let us now compare two scenarios.

Scenario 6.29. *Agent* 2 *is suspicious. When* 2 *receives the information that* 1 *has chosen* T_1 *then* 2 *believes that it is because* 1 *intends to play* T_3 *(to take the money) if he gets the opportunity. In game-theoretical terms,* 2 *believes that* 1 *will behave rationally at his last information set.*

In the event model corresponding to the preceding scenario, 2 considers L_T as strictly more likely than L_L.

Scenario 6.30. *Agent* 2 *is confident in human nature. When* 2 *receives the information that* 1 *has chosen* T_1 *then* 2 *believes that it is because* 1 *intends to play* L_3 *(to leave the money), leading to an outcome that Pareto-dominates both the* T_1*- and the* $L_1; T_2$*-outcome. In game-theoretical terms* 2 *expects that* 1 *will do something irrational at his last information set.*

In the event model corresponding to the second scenario (Figure 6.10), 2 considers L_L as strictly more likely than L_T.

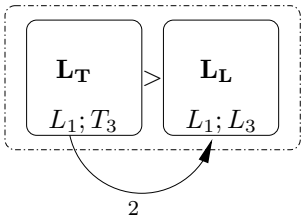

Figure 6.10: 2 is confident.

The model in Figure 6.11 displays how the informational process unfolds in the second scenario. Consider the state LLT. In the initial model, at LLT, it is indeed common belief that both players will behave rationally all the way. But when 2's initial belief that 1 will take the money immediately is defeated, she changes her mind and expects 1 to cooperate "again". So she is being rational by leaving the money, but finally 1, rationally, takes the money. This illustrates the first observation made in the literature: (initial) common belief of rationality and rationality does not imply backward induction.

If we use instead the event model corresponding to the first scenario, common belief of rationality fails. Indeed 1 would initially expect 2 to behave irrationally

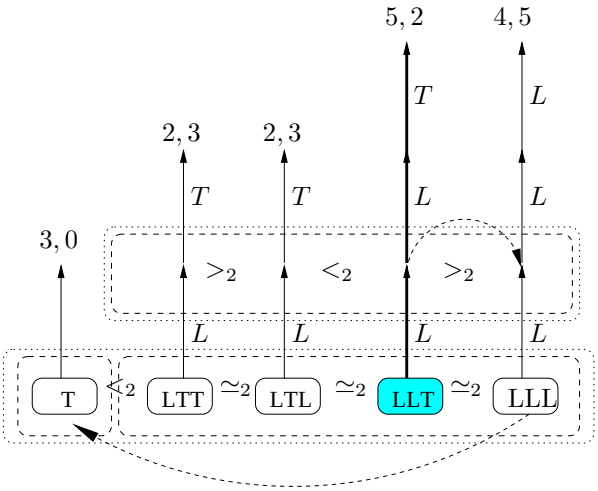

Figure 6.11: Failure of backward induction under common belief of rationality.

(and leave the money) if 1 were to leave the money in the first place. The comparison of the models illustrates the second observation made in the literature: that the epistemic foundations of backward induction might be restored by making assumptions about the way players revise their beliefs.

As we said, the preceding observations have been made before. They have also been considered from a logical perspective including in a dynamic doxastic setting in Baltag et al. [21]. But the approach presented here illustrates how using protocols and epistemic plausibility models one could capture some strategic reasoning. The natural next step is to generalize the previous dynamic approach to strategic reasoning and even more importantly to discuss how to give it a syntactic counterpart by first developing the right languages to express the important notions at work and then explore whether such languages can be axiomatized. If they can be, giving syntactic derivations of semantic arguments from the epistemic game theory literature — similarly to what we did for an agreement result when studying interactive reasoning in Chapter 4 — seems to us an interesting line of research.

Other open problems include extending our analysis of equilibrium concepts with future-oriented beliefs and studying in detail the logic of expectations initiated in Section 6.6. In general it is fairly possible that quantification over branches together with an uncertainty relation and a full temporal syntax might be enough for this logic to lack the finite model property and maybe even worse. It is anyway in our view one of the most interesting open problems in this chapter. Finally completing the picture of an analysis of the role of belief change in strategic reasoning based on protocols is another open issue. In particular giving it the right corresponding syntax and exploring the syntactic counterparts of semantic results from epistemic game theory.

Finally let us mention that Zvesper [152] offers an extensive treatment of the relation between dynamic epistemic logic and the epistemic foundations of game theory.

6.8 Conclusion

This chapter studied strategic reasoning in extensive games of imperfect (and perfect) information using approaches living at the edge of temporal and dynamic doxastic logics. We looked at solution algorithms for extensive games of imperfect information as model-change operations and extracted the corresponding notions of rationality that implicitly underly them. We initiated a doxastic temporal study of notions of equilibrium using past beliefs. We then focused on logical modeling of expectations (future-oriented beliefs) in imperfect information contexts and their properties. Finally we used our earlier framework based on protocols to discuss the role of belief change in the analysis of backward induction.

Major sources. Concerning the modeling of extensive games of imperfect information as process models the starting points were van Benthem [26], Bonanno [46] and van der Hoek and Pauly [105]. The idea of looking at solution concepts as model change operations and extracting notions of rationality corresponding to a given solution algorithm stems from van Benthem [31]. A source for our work on expectations is Bonanno [41]. In particular the assumptions on how expectations are revised are natural generalizations of the ones considered in Bonanno [42]. The idea of using a richer informational structure to model beliefs about strategies is conceptually based on Harsanyi [98] and technically in line with the perspective of Fagin et al. [72] on multi-agent systems. Finally our dynamic approach to the role of belief change in backward induction has sources in Board [40] and Baltag et al. [21], and also in Stalnaker [147, 146], Halpern [92] for the logical approach, and for the more general problematic in Reny [139] and a large literature that has shaped our way of thinking about these issues, including [111, 138, 10, 38, 25].

Our main results. This chapter analyzed strategic reasoning in extensive games of imperfect and perfect information from the perspective of a logical framework drawing on features of both temporal and dynamic logics. We have shown concretely how to develop a doxastic temporal approach based on the game structure itself, giving the starting point of a logical characterization of equilbrium concepts. We discussed a notion of expectations based on a branching-time semantics, some natural assumptions to impose on them and their syntactic counterpart. Finally we look at two ways of giving a dynamic perspective on solution concepts and their foundations. On the one hand, we discuss how rationality concepts could be extracted from solution algorithms for extensive games of imperfect information. On the other hand, we discuss how to give a protocol-based dynamic approach to the role of belief change in the foundations of backward induction.

The next step. This concludes our discussion of how to analyze strategic reasoning in extensive games of imperfect and perfect information from the perspective of doxastic temporal logics and of dynamic doxastic logics. In the next and last chapter we bring the focus from belief change to a complementary aspect of rational interaction: coalitional agency and preferences.

Chapter 7

Modal logics for preferences and cooperation: Expressivity and complexity[1]

In this dissertation we have concentrated so far on informational and doxastic aspects of interaction. In this chapter we switch the focus to another complementary dimension: that of cooperative agency and preferences, but we maintain the methodology: fixing the model, comparing the languages, chiefly in terms of their respective definability. In some sense we even systematize the methodology. But we will get into the methodological details in due time. Let us start with a brief presentation of our two fresh newcomers: cooperation and preferences.

7.1 Introduction

Cooperation of agents is a major issue in fields such as computer science, economics and philosophy. The conditions under which coalitions are formed occur in various situations involving multiple agents. A single airline company for instance cannot afford the cost of an airport runway whereas a group of companies can. Generally, agents can form groups in order to share complementary resources or because as a group they can achieve better results than individually. Modal logic (ML) frameworks for reasoning about cooperation mostly focus on what coalitions can achieve. Coalition Logic (**CL**) [131] uses modalities of the form $[\mathtt{C}]\varphi$ saying that "coalition $\mathtt{C}$ has a joint strategy to ensure that φ". **CL** has neighborhood semantics but it has been shown how it can be simulated on Kripke models [53].

Another crucial concept for reasoning about interactive situations is that of *preferences*. It has also received attention from modal logicians ([86] surveys). Recent works (e.g. [112, 1]) propose different mixtures of cooperation and prefer-

[1]This chapter is based on Dégremont and Kurzen [62].

ence logics for reasoning about cooperation. In such logics many concepts from *game theory* (GT) and *social choice theory* (SCT) are commonly encountered. Depending on the situations to be modelled, different bundles of notions are important. The ability to express these notions — together with good computational behavior — makes a logic appropriate for reasoning about the situations under consideration.

Rather than proposing a new logical framework, with specific expressivity and complexity, in this chapter, we identify how *social choice theory* and *game theory* notions are demanding for modal logics in terms of expressivity and complexity. We identify notions relevant for describing interactive situations. Some of them are local, i.e. they are properties of pointed models. We determine under which operations on models these properties are invariant. Other properties are global, i.e. they are properties of frames. For each of them, we check whether the class of frames having this property is closed under certain operations. We refer to such results as satisfiability invariance and validity closure results respectively.

We also give explicit definability results for these notions. Given a local property P we give a formula φ such that a pointed model $\mathcal{M}, w \Vdash \varphi$ iff $\mathcal{M}, w$ has property P. Given a global property Q we give a formula ψ such that a frame $\mathcal{F} \Vdash \varphi$ iff $\mathcal{F}$ has property P. We thus identify the natural (extended) modal languages needed depending on the class of frames actually considered and the particular bundle of notions of interest. We finally draw some conclusion about the complexity of reasoning about cooperation using modal logics.

Our results apply to logics interpreted on Kripke structures using a (preference) relation for each agent and a relation for each coalition. The latter can be interpreted in various ways. The pair (x, y) being in the relation for coalition $\mathtt{C}$ can e.g. mean:

- Coalition $\mathtt{C}$ considers y as being at least as good as x.
- If the system is in state x, $\mathtt{C}$ would choose y as the next state.
- $\mathtt{C}$ can submit a request such that if it is the first one received by the server while the state is in x, then the state of the system will change from x to y.
- When the system is in state x, $\mathtt{C}$ considers it possible that it is in state y.

Interpreting the relation as the possibility to bring the system into a different state applies to scenarios where agents act sequentially (e.g. with a server treating requests in a "first-come, first-served" manner) rather than simultaneously (as in ATL [4] or **CL**). In special cases — e.g. for turn-based [89, 131] frames — the approaches coincide. Still, the two approaches are first of all complementary. Our focus in this chapter is on concepts bridging powers and preferences. The same analysis is possible for powers themselves in ATL-style. Both analyses can then be combined in an interesting way. Finally, an important alternative interpretation of the coalition relation is that of group preferences, in which case ATL models

can simply be merged with the models we consider. We discuss the possible interpretations of such models in more detail in Section 7.2.

Structure of this chapter. Section 7.2 presents three classes of models of cooperative situations. Section 7.3 introduces local and global notions motivated by ideas from GT and SCT indicating local properties of a system and global properties that characterize classes of frames. Section 7.4 presents a large class of extended modal languages interpreted on the previous models. For background on invariance and closure results for modal languages, the reader can check Appendix B. In Section 7.5, we study the expressivity needed to express the local notions (to define the global properties) by giving invariance results for relevant operations and relations between models (frames). Section 7.6 completes this work by defining the notions in fragments of (extended) modal languages. We give complexity results for model checking and satisfiability for these languages and thereby give upper bounds for the complexity of logics that can express the introduced notions. Section 7.7 concludes.

7.2 The models

Our aim is to study how demanding certain concepts from game theory and social choice theory are in terms of expressivity and complexity. This depends on the models chosen. We consider three classes of simple models that have many suitable interpretations. This gives our results additional significance. A *frame* refers to the relational part of a model. For simplicity, we introduce models and assume that the domain of the valuation is a countable set of propositional letters PROP and nominals NOM. We focus on model theory and postpone discussion of formal languages to Section 7.4.

Definition 7.1 (N-LTS)**.** *A N-LTS (Labeled Transition System indexed by a finite set of agents* N*) is of the form* $\langle W, \mathtt{N}, \{ \xrightarrow{\mathtt{C}} \mid \mathtt{C} \subseteq \mathtt{N}\}, \{ \leq_j \mid j \in \mathtt{N}\}, V\rangle$*, where* $W \neq \emptyset$*,* $\mathtt{N} = \{1, \ldots, n\}$ *for some* $n \in \mathbb{N}$*,* $\xrightarrow{\mathtt{C}} \subseteq W \times W$ *for each* $\mathtt{C} \subseteq \mathtt{N}$*,* $\leq_j \subseteq W \times W$ *for each* $j \in \mathtt{N}$*, and* $V : \text{PROP} \cup \text{NOM} \to \wp(W)$*,* $|V(i)| = 1$ *for each* $i \in \text{NOM}$*.*

W is the set of states, N a set of agents and $w \xrightarrow{\mathtt{C}} v$ says that coalition C can change the state of the system from w into v. As mentioned, other interpretations are possible. $w \leq_j v$ means that j finds the state v at least as good as w. $w \in V(p)$ means that p is true at w. Preferences are usually assumed to be total pre-orders (TPO). Let TPO-N-LTS denote the class of N-LTSs in which for each $j \in \mathtt{N}$, $\leq_j$ is a TPO. We also consider models with strict preferences as explicit primitives.

Definition 7.2 (S/TPO-N-LTS)**.** *Define S/TPO-N-LTS as models of the form* $\langle W, \mathtt{N}, \{ \xrightarrow{\mathtt{C}} \mid \mathtt{C} \subseteq \mathtt{N}\}, \{ \leq_j \mid j \in \mathtt{N}\}, \{ <_j \mid j \in \mathtt{N}\}, V\rangle$*, which extend TPO-N-LTS*

models by an additional relation $<_j \subseteq W \times W$ *for each* $j \in \mathtt{N}$ *with the constraint that for each* $j \in \mathtt{N}$, $w <_j v$ *iff* $w \leq_j v$ and $v \not\leq_j w$.

Depending on the interpretation of $\xrightarrow{\mathtt{C}}$, it can be complemented or replaced by effectivity functions (**CL**) or more generally transition functions as in ATL. In the latter sense, powers of coalitions will in general not reduce to relations on states. We leave an analysis of powers in such settings aside for now. There would be two ways to go: drawing on the model-theory of neighborhood semantics [95] or on a normal simulation of **CL** [53]. Generally, the expressive power might depend on whether coalitional powers are taken as primitives or computed from individual powers.

In the next section, we identify a list of notions inspired by concepts from game theory and social choice theory for reasoning about cooperative ability and preferences of agents. Then we will determine the expressivity required by certain local and global notions, by giving invariance results for pointed models and closure conditions for classes of frames, respectively. Since we are also interested in the effects of the underlying models on the expressivity required to express the local notions, we will give invariance results with respect to the three different types of models we just introduced.

7.3 The notions

Reasoning about cooperative interaction considers what coalitions of agents can achieve and what individuals prefer. Using these elements, more elaborate notions can be built. We consider natural counterparts of SCT and GT notions and are interested both in local notions i.e. properties of a particular state in a particular system, i.e. properties of pointed models $\mathcal{M}, w$, and also in global notions, which are properties of classes of systems. In other words, we are interested in the class of frames a global property characterizes. With respect to content, apart from notions describing only coalitional powers or preferences, we consider stability and effectivity concepts.

Power of Coalitions. We now present some interesting notions about coalitional power. Recall that $w \xrightarrow{\mathtt{C}} v$ can e.g. mean "$\mathtt{C}$ can achieve v at w".

Local Notions. Interesting properties of coalitional power involve the relation between the powers of different groups (*PowL*3) and the contribution of individuals to a group's power, e.g. an agent is needed to achieve something (*PowL*2).

- *PowL*1. Coalition $\mathtt{C}$ can achieve a state where p is true. $\exists x(w \xrightarrow{C} x \wedge P(x))$
- *PowL*2. Only groups with i can achieve p-states. $\bigwedge_{\mathtt{C} \subseteq \mathtt{N} \setminus i}(\forall x(w \xrightarrow{\mathtt{C}} x \Rightarrow \neg P(x)))$
- *PowL*3. Coalition $\mathtt{C}$ can force every state that coalition $\mathtt{D}$ can force. $\forall x(w \xrightarrow{\mathtt{D}} x \Rightarrow w \xrightarrow{\mathtt{C}} x)$

Global Notions. *PowG*1 says that each coalition can achieve exactly one result. *PowG*3 expresses coalition monotonicity: it says that if a coalition can achieve some result, then so can every superset of that coalition. In many situations, decision making in groups can only be achieved by a majority (*PowG*2). *PowG*4 and *PowG*5 exemplify (mathematically natural) consistency requirements between powers of non-overlapping coalitions.

- *PowG*1. In any state each coalition can achieve exactly one state.
  $\bigwedge_{C\subseteq \mathtt{N}} \forall x \exists y (x \xrightarrow{C} y \wedge \forall z (x \xrightarrow{C} z \Rightarrow z = y))$
- *PowG*2. Only coalitions containing a majority of $\mathtt{N}$ can achieve something.
  $\forall x (\bigwedge_{\mathtt{C}\subseteq \mathtt{N}, |\mathtt{C}| < \frac{|N|}{2}} (\neg \exists y (x \xrightarrow{\mathtt{C}} y)))$
- *PowG*3. Coalition monotonicity, i.e. if for $\mathtt{C}$ and $\mathtt{D}$, $\mathtt{C} \subseteq \mathtt{D}$, then $R_{\mathtt{C}} \subseteq R_{\mathtt{D}}$.
  $\forall x (\bigwedge_{\mathtt{C}\subseteq \mathtt{N}} \bigwedge_{\mathtt{D}\subseteq \mathtt{N}, \mathtt{C}\subseteq \mathtt{D}} (\forall y (x \xrightarrow{\mathtt{C}} y \Rightarrow x \xrightarrow{\mathtt{D}} y)))$
- *PowG*4. If $\mathtt{C}$ can achieve something, then subsets of its complement cannot achieve anything.
  $\forall x \bigwedge_{\mathtt{C}\subseteq \mathtt{N}} (((\exists y (x \xrightarrow{\mathtt{C}} y)) \Rightarrow \bigwedge_{\mathtt{D}\subseteq \mathtt{N}\setminus \mathtt{C}} \neg \exists z (x \xrightarrow{\mathtt{D}} z)))$
- *PowG*5. If $\mathtt{C}$ can achieve something, then subsets of its complement cannot achieve something $\mathtt{C}$ cannot achieve.
  $\forall x \bigwedge_{\mathtt{C}\subseteq \mathtt{N}} (((\exists y (x \xrightarrow{\mathtt{C}} y)) \Rightarrow \bigwedge_{\mathtt{D}\subseteq \mathtt{N}\setminus \mathtt{C}} \forall z (x \xrightarrow{\mathtt{D}} z \Rightarrow x \xrightarrow{\mathtt{C}} z)))$

Preferences. What do agents prefer? What are suitable global constraints on preferences? $w \leq_i v$ means "*i* finds *v at least as good* (a.l.a.g.) as *w*". We write $w <_i v$ for $w \leq_i v \wedge \neg(v \leq_i w)$, meaning that "*i strictly prefers v* over *w*".

Local Notions. First of all, we can distinguish between strict and nonstrict preferences. The most basic preference relation that we consider is that of being a.l.a.g. We can also look at the relation "at least as bad" (a.l.a.b) (*PrefL*4). Agents' preferences over states can also be seen as being based on preferences over propositions [60]. *PrefL*8 (*PrefL*10) says the truth of a given proposition is a sufficient (necessary) condition for an agent to prefer some state. In what follows, "at least as good" (a.l.a.g) means "at least as good *as the current state*".

- *PrefL*1. There is a state *i* finds a.l.a.g. where *p* holds. $\exists x (w \leq_i x \wedge P(x))$
- *PrefL*2. There is a *p*-state that *i* strictly prefers. $\exists x (w <_i x \wedge P(x))$
- *PrefL*3. There is a state that all agents find a.l.a.g and that at least one strictly prefers. $\exists x (\bigwedge_{i\in \mathtt{N}} (w \leq_i x) \wedge \bigvee_{j\in \mathtt{N}} w <_j x)$
- *PrefL*4. There is a state that *i* finds a.l.a.b. where *p* holds.
  $\exists x (x \leq_i w \wedge P(x))$

- *PrefL*5. There is a state that i finds strictly worse where p is true.
  $\exists x(x <_i w \wedge P(x))$
- *PrefL*6. i finds a state a.l.a.g. a the current one iff j does.
  $\forall x(w \leq_i x \leftrightarrow w \leq_j x)$
- *PrefL*7. There is a state only i finds a.l.a.g. $\exists x(w \leq_i x \wedge \bigwedge_{j \in \mathbb{N} \setminus \{i\}} \neg(w \leq_j x))$
- *PrefL*8. i finds every p-state a.l.a.g. $\forall x(P(x) \Rightarrow w \leq_i x)$
- *PrefL*9. i strictly prefers every p-state. $\forall x(P(x) \Rightarrow w <_i x)$
- *PrefL*10. i considers only p-states to be a.l.a.g. $\forall x(w \leq_i x \Rightarrow P(x))$
- *PrefL*11. i strictly prefers only p-states. $\forall x(w <_i x \Rightarrow P(x))$

Global Notions. Capturing the intuitive idea of preferences requires several conditions for the preference relation: reflexivity, transitivity and completeness (trichotomy for strict preferences). Sometimes, it can also be appropriate to say that for each alternative there is exactly one that is at least as good (*PrefG*8).

- *PrefG*1. "at least as good as" is reflexive. $\forall x(\bigwedge_{i \in \mathbb{N}}(x \leq_i x))$
- *PrefG*2. "at least as good as" is transitive.
  $\forall x \forall y \forall z(\bigwedge_{i \in \mathbb{N}}((x \leq_i y \wedge y \leq_i z) \Rightarrow x \leq_i z))$
- *PrefG*3. "at least as good as" is complete.
  $\forall x \forall y(\bigwedge_{i \in \mathbb{N}}(x \leq_i y \vee y \leq_i x))$
- *PrefG*4. "at least as good as" is a total pre-order.
  (Conjunction of the two previous formulas.)
- *PrefG*5. "strictly better than" is transitive.
  $\forall x \forall y \forall z((\bigwedge_{i \in \mathbb{N}}(x <_i y \wedge y <_i z) \Rightarrow x <_i z)))$
- *PrefG*6. "strictly better than" is trichotomous.
  $\forall x \forall y(\bigwedge_{i \in \mathbb{N}}(x <_i y \vee y <_i x \vee x = y))$
- *PrefG*7. "strictly better than" is a strict total order.
  (Conjunction of the previous two formulas.)
- *PrefG*8. Determinacy for "at least as good as", i.e. exactly one successor.
  $\forall x(\bigwedge_{i \in \mathbb{N}}(\exists y(w \leq_i y \wedge \forall z(x \leq_i z \Rightarrow z = y))))$

So far, we have focused on preferences of individuals. A natural question in SCT is how to aggregate individual preferences into group preferences. We can address this question by interpreting $\xrightarrow{\mathtt{C}}$ as a preference relation for each $\mathtt{C} \subseteq \mathtt{N}$.

- *PrefG*9. C finds a state a.l.a.g. as the current one iff all its members do.
  $\forall x \forall y(\bigwedge_{\mathtt{C}\subseteq\mathtt{N}}(x \xrightarrow{\mathtt{C}} y \leftrightarrow \bigwedge_{i\in\mathtt{C}} x \leq_i y))$
- *PrefG*10. C finds a state at least as good as the current one iff at least one member does.
  $\forall x \forall y(\bigwedge_{\mathtt{C}\subseteq\mathtt{N}}(x \xrightarrow{\mathtt{C}} y \leftrightarrow \bigvee_{i\in\mathtt{C}} x \leq_i y))$
- *PrefG*11. C finds a state a.l.a.g. as the current one iff most members do.
  $\forall x \forall y(\bigwedge_{\mathtt{C}\subseteq\mathtt{N}}(x \xrightarrow{\mathtt{C}} y \leftrightarrow \bigvee_{\mathtt{D}\subseteq\mathtt{C},|\mathtt{D}|>\frac{|\mathtt{C}|}{2}}(\bigwedge_{i\in\mathtt{D}} x \leq_i y)))$

Combining the preceding concepts. We start with the conceptually and historically important SCT notion of a *dictator*. d is a dictator if the group's preferences mimic d's preferences. Interpreting $\xrightarrow{\mathtt{C}}$ as an achievement relation, we get an even stronger notion: groups can only *do* what d likes. A *local* dictator is a dictator who controls one state in the system, and a *dictator* controls all states.

Definition 7.3 (Local Dictatorship)**.** *i is a* weak (strong) local dictator *at w iff any group prefers v at w only if for i, v is a.l.a.g. as (strictly better than) w.*

We now introduce combinations of powers and preferences. The first notion says that coalition C can do something useful for i (in some cases giving i an incentive to join) and the third notion characterizes situations in which a unanimously desired state remains unachievable. We start with **Local Notions**.

- *PPL*1. C can achieve a state that i finds at least as good as the current one.
  $\exists x(w \xrightarrow{\mathtt{C}} x \wedge w \leq_i x)$
- *PPL*2. C can achieve a state that all $i \in \mathtt{D}$ find a.l.a.g. as the current one.
  $\exists x(w \xrightarrow{\mathtt{C}} x \wedge \bigwedge_{i\in\mathtt{D}} w \leq_i x)$
- *PPL*3. There is a state that all agents prefer but no coalition can achieve it. $\exists x((\bigwedge_{i\in\mathtt{N}} w \leq_i x) \wedge \bigwedge_{\mathtt{C}\subseteq\mathtt{N}} \neg(w \xrightarrow{\mathtt{C}} x))$
- *PPL*4. C can achieve all states that agent i finds a.l.a.g. as the current one.
  $\forall x(w \leq_i x \Rightarrow w \xrightarrow{\mathtt{C}} x)$
- *PPL*5. C can achieve all states that i strictly prefers. $\forall x(w <_i x \Rightarrow w \xrightarrow{\mathtt{C}} x)$
- *PPL*6. i is a weak local dictator. $\forall x(\bigwedge_{\mathtt{C}\subseteq\mathtt{N}}(w \xrightarrow{\mathtt{C}} x \Rightarrow w \leq_i x))$

- *PPL*7. i is a strong local dictator. $\forall x(\bigwedge_{\mathtt{C}\subseteq\mathtt{N}}(w \xrightarrow{\mathtt{C}} x \Rightarrow w <_i x))$

Global Notions. *PPG*1 is a natural constraint on coalitional power: a group can achieve a state iff it is good for all members — otherwise they would not take part in the collective action. *PPG*3 is a condition of Arrow's impossibility theorem. *PPG*4 reflects individual rationality: don't join a group if you don't gain anything. It can be generalized to every sub-coalition or weakened to "not joining if you lose something" (cf. the core of a coalitional game [127] (Def. 268.3)). *PPG*5 applies to systems where an agent is indispensable to achieve anything: a unique capitalist in a production economy or a unique server are typical examples.

- *PPG*1. Coalitions can only achieve states that all their members consider at least as good as the current one. $\forall x\forall y \bigwedge_{\mathtt{C}\subseteq\mathtt{N}}(x \xrightarrow{\mathtt{C}} y \Rightarrow \bigwedge_{i\in\mathtt{C}}(x \leq_i y))$
- *PPG*2. One agent is a weak local dictator in every state (*dictator*). $\bigvee_{i\in\mathtt{N}} \forall x\forall y(x \xrightarrow{\mathtt{C}} y \Rightarrow x \leq_i y)$
- *PPG*3. There is no *dictator*. $\neg(\bigvee_{i\in\mathtt{N}} \forall x\forall y(x \xrightarrow{\mathtt{C}} y \Rightarrow x \leq_i y))$
- *PPG*4. If i can achieve some state i strictly prefers then for any $\mathtt{C}$ containing i: if $\mathtt{C} \setminus i$ cannot achieve some state but $\mathtt{C}$ can, then i strictly prefers that state. $\bigwedge_{i\in\mathtt{N}} \forall x(\exists y(x \xrightarrow{\{i\}} y \wedge x <_i y) \Rightarrow \bigwedge_{\mathtt{C}\subseteq\mathtt{N}, i\in\mathtt{C}}(\forall z(x \xrightarrow{\mathtt{C}} z \wedge \neg(x \xrightarrow{\mathtt{C}\setminus\{i\}} z)) \Rightarrow x <_i z))$
- *PPG*5. Only groups with i can achieve something. $\forall x \bigwedge_{\mathtt{C}\subseteq\mathtt{N}\setminus\{i\}} \neg\exists y(x \xrightarrow{\mathtt{C}} y)$
- *PPG*6. In all states, there is an i such that groups with i can achieve exactly the states as they can without i. $\forall x(\bigvee_{i\in\mathtt{N}} \bigwedge_{\mathtt{C}\subseteq\mathtt{N}, i\in\mathtt{C}} \forall y(x \xrightarrow{\mathtt{C}} y \leftrightarrow x \xrightarrow{\mathtt{C}\setminus\{i\}} y))$
- *PPG*7. For any agent, there is some state in which coalitions not containing this agent cannot achieve any state. $\bigwedge_{i\in\mathtt{N}} \exists x(\bigwedge_{\mathtt{C}\subseteq\mathtt{N}, i\in\mathtt{C}} \neg\exists y(x \xrightarrow{\mathtt{C}} y))$

Efficiency and Stability Notions. In our setting, it is natural to interpret the state space as possible social states or allocations of goods. A criterion from welfare economics to distinguish "good" from "bad" states is that of *efficiency*: if we can change the allocation or social state and make an agent happier without making anyone less happy then we are using resources more efficiently and it is socially desirable to do so. E.g. *PrefL*3 in this respect means that the current state is not efficient: there is a state that is a *Pareto-improvement* of it. Importing the notion of Pareto-efficiency into our framework is straightforward.

Definition 7.4 (Pareto-efficiency)**.** *A state is weakly (strongly)* Pareto-*efficient iff there is no state that everyone strictly prefers (finds a.l.a.g). A state is* Pareto-*efficient iff there is no state such that everyone considers it to be at least as good and at least one agent thinks that it is strictly better.*

GT equilibrium concepts characterize stable states: given what others are doing, I don't have an incentive to do something that makes us leave this stable state. Generalizing, a system is in a stable state if nobody has an incentive to change its current state. We can think of strategy profiles in a strategic game as assigning roles to the agents. Two profiles $x = (s^*_{-i}, s^*_i), y = (s^*_{-i}, s'_i)$ are related by $\overset{\{i\}}{\rightarrow}$ iff i can unilaterally change role (strategy) to s'_i in the next round of the game. E.g. the stability of a state where an agent provides the public good on his own depends on whether he cares enough about it to provide it on his own. A state is stable iff there is no strictly preferred state that an agent can achieve alone. Since the idea relates to *Nash* equilibria (see [127]), we use the names *Nash-stability*, and *Nash-cooperation stability* for its group version.

Definition 7.5 (Nash-stability)**.** *A state is (strongly)* Nash*-stable iff there is no state that an agent i strictly prefers (finds a.l.a.g.) and that i can achieve alone. It is (strongly)* Nash*-cooperation stable iff there is no state v and coalition $\mathtt{C}$ such that every $i \in \mathtt{C}$ strictly prefers v (finds v a.l.a.g.) and $\mathtt{C}$ can achieve v.*

Local Notions

- *EF*1. The current state is weakly *Pareto*-efficient. $\neg\exists x(\bigwedge_{i\in\mathtt{N}}(w <_i x))$
- *EF*2. The current state is *Pareto*-efficient. $\neg\exists x((\bigwedge_{i\in\mathtt{N}} w \leq_i x) \wedge \bigvee_{j\in\mathtt{N}} w <_i x)$
- *EF*3. The current state is strongly *Pareto*-efficient. $\neg\exists x(\bigwedge_{i\in\mathtt{N}} w \leq_i x)$
- *ST*1. The current state is *Nash* stable. $\neg\exists x(\bigvee_{i\in\mathtt{N}}(w \overset{\{i\}}{\rightarrow} x \wedge w <_i x))$
- *ST*2. The current state is strongly *Nash* stable. $\neg\exists x(\bigvee_{i\in\mathtt{N}}(w \overset{\{i\}}{\rightarrow} x \wedge w \leq_i x))$
- *ST*3. The current state is strongly is *Nash*-cooperation stable. $\neg\exists x(\bigvee_{C\subseteq\mathtt{N}}(w \overset{\mathtt{C}}{\rightarrow} x \wedge \bigwedge_{i\in\mathtt{C}} w <_i x))$
- *ST*4. The current state is strongly *Nash*-cooperation stable. $\neg\exists x(\bigvee_{C\subseteq\mathtt{N}}(w \overset{\mathtt{C}}{\rightarrow} x \wedge \bigwedge_{i\in\mathtt{C}} w \leq_i x))$

This concludes the list of notions we will consider. We now turn to the possible extended modal languages that can express them, interpreting them on the models of Section 7.2, and making observations on their expressive power.

7.4 Modal languages and their expressivity

As will be clear from the invariance results of the next sections, the basic modal language will generally be too weak for reasoning about cooperation. However,

any notion expressible in the FO correspondence language is expressible in the hybrid language $\mathcal{H}(\mathtt{E}, @, \downarrow)$ [55]. We have introduced the ideas behind some extended modal languages in Section 1.7 such as hybrid logics and boolean modal logics. Among them are model-theoretically well-understood fragments. We introduce all these **Extended Modal Languages** at once as a "super" logic interpreted on our earlier models. For background on invariance and closure results for these logics the reader is referred to Appendix B.

Syntax. The syntax of this "super" logic is defined by simultaneous recursion as follows:

$$\alpha ::= \leq_j \mid \mathtt{C} \mid \alpha^{-1} \mid ?\varphi \mid \alpha;\alpha \mid \alpha \cup \alpha \mid \alpha \cap \alpha \mid \overline{\alpha}$$

$$\varphi ::= p \mid i \mid x \mid \neg\varphi \mid \varphi \wedge \varphi \mid \langle\alpha\rangle\varphi \mid \mathtt{E}\varphi \mid @_i\varphi \mid @_x\varphi \mid \downarrow x.\varphi \mid [\![\alpha]\!]\varphi$$

where $j \in \mathtt{N}$, $\mathtt{C} \in \wp(\mathtt{N}) - \{\emptyset\}$, p ranges over PROP, i ranges over NOM and $x \in$ SVAR, for SVAR being a countable set of variables.

Semantics. A valuation maps propositional letters to subsets of the domain and nominals to singleton subsets. Given a N-LTS, a program α is interpreted as a relation as follows:

$$\begin{array}{lcl} R_{\leq_i} & = & \leq_i \\ R_C & = & \stackrel{C}{\rightarrow} \\ R_{\beta^{-1}} & = & \{(v,w) | wR_\beta v\} \\ R_{\beta\cup\gamma} & = & R_\beta \cup R_\gamma \\ R_{\beta\cap\gamma} & = & R_\beta \cap R_\gamma \\ R_{\overline{\beta}} & = & (W \times W) - R_\beta \end{array}$$

Formulas are interpreted together with an assignment $g : \text{SVAR} \to W$ as indicated below. We skip booleans.

$$\begin{array}{lll} \mathcal{M}, w, g \Vdash i & \text{iff} & w \in V(i) \\ \mathcal{M}, w, g \Vdash x & \text{iff} & w = g(x) \\ \mathcal{M}, w, g \Vdash \langle\alpha\rangle\varphi & \text{iff} & \text{for some } v \text{ with } wR_\alpha v \text{ we have } \mathcal{M}, v, g \Vdash \varphi \\ \mathcal{M}, w, g, \Vdash \mathtt{E}\varphi & \text{iff} & \text{for some } v \in W \text{ we have } \mathcal{M}, v, g \Vdash \varphi \\ \mathcal{M}, w, g, \Vdash @_i\varphi & \text{iff} & \mathcal{M}, v, g \Vdash \varphi \text{ where } V(i) = \{v\} \\ \mathcal{M}, w, g, \Vdash @_x\varphi & \text{iff} & \mathcal{M}, g(x), g \Vdash \varphi \\ \mathcal{M}, w, g, \Vdash \downarrow x.\varphi & \text{iff} & \mathcal{M}, w, g[x := w] \Vdash \varphi \\ \mathcal{M}, w, g \Vdash [\![\alpha]\!]\varphi & \text{iff} & \text{for all } w \in W \text{ we have } wR_\alpha v \text{ whenever } \mathcal{M}, v, g \Vdash \varphi \end{array}$$

Expressivity. The least expressive modal language we consider is $\mathcal{L}(\mathtt{N})$, which is of similarity type $\langle(\mathtt{C})_{\mathtt{C}\subseteq\mathtt{N}}, (\leq_i)_{i\in\mathtt{N}}\rangle$. Its natural extensions go along two lines: adding program constructs and new operators. $\mathcal{L}(\mathtt{N}, \cap, i)$ e.g. refers to the logic with language: $\alpha ::= \leq_\mathtt{j} \mid \mathtt{C} \mid \alpha \cap \alpha \quad \varphi ::= p \mid i \mid \neg\varphi \mid \varphi \wedge \varphi \mid \langle\alpha\rangle\varphi$. As language inclusion implies expressivity inclusion (indicated by "$\leq$"), we only indicate (some) non-obvious facts of inclusions in this space of modal languages.

Fact 7.6. $\mathcal{L}(\mathtt{N}, \cup, \ ; \ , ?) \leq \mathcal{L}(\mathtt{N})$.

Proof. By the facts that $\mathcal{M}, w, g \Vdash \langle \alpha \cup \beta \rangle \varphi$ iff $\mathcal{M}, w, g \Vdash \langle \alpha \rangle \varphi \vee \langle \beta \rangle \varphi$; that $\mathcal{M}, w, g \Vdash \langle \alpha ; \beta \rangle \varphi$ iff $\mathcal{M}, w, g \Vdash \langle \alpha \rangle \langle \beta \rangle \varphi$ and that $\mathcal{M}, w, g \Vdash \langle ?\psi \rangle \varphi$ iff $\mathcal{M}, w, g \Vdash \psi \wedge \varphi$. QED

Fact 7.7. $\mathcal{L}(\mathtt{N}, @, i) \leq \mathcal{L}(\mathtt{N}, \mathtt{E}, i)$.

Proof. By the fact that $\mathcal{M}, w, g \Vdash @_i \varphi$ iff $\mathcal{M}, w, g \Vdash \mathtt{E}(i \wedge \varphi)$. QED

Fact 7.8. $\mathcal{L}(\mathtt{N}, \cap) \leq \mathcal{L}(\mathtt{N}, \downarrow, @, x)$.

Proof. $\mathcal{M}, w, g \Vdash \langle \alpha \cap \beta \rangle \varphi$ iff $\mathcal{M}, w, g \Vdash \downarrow x. \langle \alpha \rangle \downarrow y. (\varphi \wedge @_x \langle \beta \rangle y)$. QED

Fact 7.9. $\mathcal{L}(\mathtt{N}, [\![\ ]\!]) \leq \mathcal{L}(\mathtt{N}, \_)$.

Proof. By the fact that $\mathcal{M}, w, g \Vdash [\![\alpha]\!] \varphi$ iff $\mathcal{M}, w, g \Vdash [\overline{\alpha}] \neg \varphi$. QED

Fact 7.10. $\mathcal{L}(\mathtt{N}, \_) \leq \mathcal{L}(\mathtt{N}, \downarrow, \mathtt{E}, x)$.

Proof. $\mathcal{M}, w, g \Vdash \langle \overline{\alpha} \rangle \varphi$ iff $\mathcal{M}, w, g \Vdash \downarrow x. \mathtt{E} \downarrow y. (\varphi \wedge \neg \mathtt{E}(x \wedge \langle \alpha \rangle y))$. QED

Fact 7.11. $\mathcal{L}(\mathtt{N}, \ ^{-1}) \leq \mathcal{L}(\mathtt{N}, \downarrow, \mathtt{E}, x)$.

Proof. By the fact that $\mathcal{M}, w, g \Vdash \langle \alpha^{-1} \rangle \varphi$ iff $\mathcal{M}, w, g \Vdash \downarrow x. \mathtt{E}(\varphi \wedge \langle \alpha \rangle x)$. QED

Fact 7.12. $\mathcal{L}(\mathtt{N}, \mathtt{E}) \leq \mathcal{L}(\mathtt{N}, \_)$.

Proof. By the fact that $\mathcal{M}, w, g \Vdash \mathtt{E}\varphi$ iff $\mathcal{M}, w, g \Vdash \langle \alpha \rangle \varphi \vee \langle \overline{\alpha} \rangle \varphi$. QED

The reader might now like to see immediately how the notions can be defined in extended modal languages and go directly to Sect. 7.6. Of course, the choice of the languages is only justified once we have determined the required expressive power both to express the local notions and to define the class of frames corresponding to the global ones. Thus we start by doing so in the next section.

7.5 Invariance and closure results

We start with satisfiability invariance results for the classes of pointed models defined in Section 7.2. Then we turn to closure results for classes of frames defined by global notions. A "Y" in a cell means that the row notion is invariant under the column operation. The numbers in the columns refer to representative proofs for these results found below the tables. They will give the reader a concrete idea of the meaning of these results. The particular choice of operations we consider is naturally determined by background invariance and closure results for modal languages (see Appendix B for details).

Overview of the Results for the General Case.

	Bis	CBis	∩-Bis	TBis	$\mathcal{H}$-Bis	$\mathcal{H}$(@)-Bis	$\mathcal{H}$(E)-Bis	BM	GSM	DU
[*PowL*1]	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
[*PowL*2]	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
[*PowL*3]	N	N	N	N	N	N	N	N	Y	Y
[*PrefL*1]	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
[*PrefL*2]	N	N	N	N	N	N	N	N	Y	Y
[*PrefL*3]	N	N	N	N	N	N	N	N	Y	Y
[*PrefL*4]	N	Y	N	N	N	N	N	N	N (14)	Y
[*PrefL*5]	N	N	N	N	N	N	N	N	N	Y
[*PrefL*6]	N	N	N	N	N	N	N	N	Y	Y
[*PrefL*7]	N	N	N	N	N	N	N	N	Y	Y
[*PrefL*8]	N	N	N	N	N	N	N	N	N	N
[*PrefL*9]	N	N	N	N	N	N	N	N	N	N
[*PrefL*10]	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
[*PrefL*11]	N	N	N	N	N	N	N	N	Y	Y
[*PPL*1]	N	N	Y	N	N	N	N	N	Y	Y
[*PPL*2]	N	N	Y	N	N	N	N	N	Y	Y
[*PPL*3]	N	N	N	N	N	N	N	N	Y	Y
[*PPL*4]	N	N	N	N	N	N	N	N	Y	Y
[*PPL*5]	N	N	N	N	N	N	N	N	Y	Y
[*PPL*6]	N	N	N	N	N	N	N	N	Y	Y
[*PPL*7]	N	N	N	N	N	N	N	N	Y	Y
[*EF*1]	N	N	N	N	N	N	N	N	Y	Y
[*EF*2]	N	N	N	N	N	N	N	N	Y	Y
[*EF*3]	N	N	Y	N	N	N	N	N	Y	Y
[*ST*1]	N	N	N	N	N	N	N	N	Y	Y
[*ST*2]	N	N	Y	N	N	N	N	N	Y	Y
[*ST*3]	N	N	N (13)	N	N	N	N	N	Y	Y
[*ST*4]	N	N	Y	N	N	N	N	N	Y	Y

Comments. Most of our notions are not bisimulation-invariant. The basic modal language of similarity type $\langle \{ \xrightarrow{\mathtt{C}} \mid \mathtt{C} \subseteq \mathtt{N}\}, \{ \leq_i \mid i \in \mathtt{N}\}\rangle$ is thus not expressive enough to describe our local notions (without restrictions on the class of frames). Invariance under BM often fails; some failures are due to intersections of relations, but as ∩-Bis also fails often, this cannot be the only reason. By contrast, invariance under GSM generally holds; it fails for properties with backward looking features. This is good news for expressivity: we can expect definability in the hybrid language with ↓-binder.[2] But not for computability, since the satisfiability problem of the bounded fragment is undecidable. Finally, the results are the same

[2]Indeed, [73, 9] have proven that all notions definable in the first-order correspondence language that are invariant under GSM are equivalent to a formula of the bounded fragment, i.e. of the hybrid language with ↓-binder (which are notational variants).

for hybrid and basic bisimulations. No surprise: roughly speaking, at the level of local satisfaction, to exploit the expressive power of nominals, the notions would have to refer explicitly to some state. Here are two representative results.

Representative Proofs for the General Case

Proposition 7.13. *On* `N-LTS`*, ST*3 *is not invariant under* $\cap$*-bisimulation.*

Proof. Let $\mathcal{M} = \langle\{w, v\}, \{1, 2\}, \{\stackrel{\mathtt{C}}{\rightarrow} | \mathtt{C} \subseteq \{1, 2\}\}, \{\leq_1, \leq_2\}, V\rangle$, where $w \stackrel{\{1,2\}}{\rightarrow} v, w \leq_1 v, v \leq_1 w, w \leq_2 v, V(p) = \{w, v\}$. Let $\mathcal{M}' = \langle\{s, t, u\}, \{1, 2\}, \{\stackrel{\mathtt{C}}{\rightarrow}' | \mathtt{C} \subseteq \{1, 2\}\}, \{\leq'_1, \leq'_2\}, V'\rangle$, where $s \stackrel{\{1,2\}}{\rightarrow}' t, u \stackrel{\{1,2\}}{\rightarrow}' t, s \leq'_1 t, u \leq'_1 t, t \leq'_1 u, s \leq'_2 t, u \leq'_2 t, V'(p) = \{s, t, u\}$. Then, $\mathcal{M}, w \Vdash ST3$ and $\mathcal{M}', s \nVdash ST3$ because $s \stackrel{\{1,2\}}{\rightarrow}' t$ and $s <'_1 t, s <'_2 t$. Moreover, $Z = \{(w, s), (w, u), (v, t)\}$ is a $\cap$-bisimulation. QED

Proposition 7.14. *On* `N-LTS`*, PrefL*4 *is not invariant under* `GSM`*.*

Proof. Let $\mathcal{M} = \langle\{w, v\}, \{1\}, \{\stackrel{\mathtt{C}}{\rightarrow} | \mathtt{C} \subseteq \{1\}\}, \{\leq_1\}, V\rangle$, where $\stackrel{\{1\}}{\rightarrow} = \emptyset$, $v \leq_1 w, V(p) = \{v\}$. Then, $\mathcal{M}, w \Vdash$ *PrefL*4 because $v \leq_1 w$ and $v \in V(p)$. But for the submodel $\mathcal{M}'$ generated by $\{w\}$, we have $\mathcal{M}', w \nVdash$ *PrefL*4 since v is not contained in $\mathcal{M}'$. QED

Results Overview for the Total Pre-orders (`TPO`) Case. This table shows rows that differ from the general case. Entries that differ are in boldface.

	Bis	CBis	$\cap$-Bis	TBis	$\mathcal{H}$-Bis	$\mathcal{H}$(@)-Bis	$\mathcal{H}$(E)-Bis	BM	GSM	DU
[*PrefL*8]	N	N	N	N	N	N	N	N	N	**Y***
[*PrefL*9]	N	**Y** (15)	N	N	N	N	N	N	N	**Y***
[*EF*1]	N	N	N	N	N	N	N	N	Y	**Y***

Proposition 7.15. *On* `TPO-N-LTS`*, PrefL*9 *is invariant under CBis.*

Proof. Let $\mathcal{M}, w$ and $\mathcal{M}', w'$ be two pointed `TPO-N-LTS` such that there exists a C-Bisimulation Z between $\mathcal{M}$ and $\mathcal{M}'$ such that $(w, w') \in Z$. But now assume that $\mathcal{M}, w \nVdash$ *PrefL*9. It follows that there exists some $t \in W$ such that $t \in V^{\mathcal{M}}(p)$ but $w \not\leq_i^{\mathcal{M}} t$. But then by totality of $\leq_i^{\mathcal{M}}$ we have $t \leq_i^{\mathcal{M}} w$, i.e. $w \geq_i^{\mathcal{M}} t$. But then by C-Bisimulation there exists some state $t' \in W'$ such that $(t, t') \in Z$ and therefore $t' \in V^{\mathcal{M}'}(p)$ and $w' \geq_i^{\mathcal{M}'} t'$ and thus $w' \not\leq_i^{\mathcal{M}'} t'$. By definition we have thus $\mathcal{M}', w' \nVdash$ *PrefL*9. The other direction is symmetrical. QED

Comments. Except for disjoint union (`DU`), the restriction to the `TPO` case brings only slight benefits. The * marks trivial invariance: the only `DU` of models that is complete is the trivial one: mapping a model to itself.

Overview of the Results for the `TPO` Case with Strict Preferences. The following table contains the rows that differ from the ones in the table for total preorders without a strict preference relation.

	Bis	CBis	∩-Bis	TBis	$\mathcal{H}$-Bis	$\mathcal{H}(@)$-Bis	$\mathcal{H}(\mathbf{E})$-Bis	BM	GSM	DU
[*PrefL*2]	**Y**	**Y**	**Y**	**Y**	**Y**	**Y**	**Y**	**Y**	Y	Y
[*PrefL*3]	N	N	N	N	N	N	N	**Y**	Y	Y
[*PrefL*5]	N	**Y**	N	N	N	N	N	N	N	Y
[*PrefL*6]	N	N	N	N	N	N	N	**Y**	Y	Y
[*PrefL*7]	N	N	N	N	N	N	N	**Y**	Y	Y
[*PrefL*8]	N	**Y**	N	N	N	N	N	N	N	Y
[*PrefL*11]	**Y**	**Y**	**Y**	**Y**	**Y**	**Y**	**Y**	**Y** (16)	Y	Y
[*PPL*7]	N	N	N	N	N	N	N	**Y**	Y	Y
[*EF*1]	N	N	**Y**	N	N	N	N	**Y**	Y	Y
[*EF*2]	N	N	**Y**	N	N	N	N	**Y**	Y	Y
[*ST*1]	N	N	**Y**	N	N	N	N	**Y** (17)	Y	Y
[*ST*3]	N	N	**Y**	N	N	N	N	**Y**	Y	Y

Proposition 7.16. *On S/TPO-N-LTS, PrefL*11 *is invariant under* BM.

Proof. Let $\mathcal{M}$ and $\mathcal{M}'$ be two S/TPO-N-LTS and assume that f is a bounded morphism from $\mathcal{M}$ to $\mathcal{M}'$. Assume that the property *PrefL*11 does not hold for $\mathcal{M}, w$, i.e. there is a state $v \in Dom(\mathcal{M})$ such that $w <_i^{\mathcal{M}} v$ and $v \notin V^{\mathcal{M}}(p)$. But then by R-homomorphism, we have $f(w) <_i^{\mathcal{M}'} f(v)$ and by Atomic Harmony, $f(v) \notin V^{\mathcal{M}'}(p)$, and thus *PrefL*11 does not hold for $\mathcal{M}', f(w)$. For the other direction, assume that *PrefL*11 is not satisfied at $\mathcal{M}', f(w)$. It follows that there is a state $v' \in Dom(\mathcal{M}')$ such that $f(w) <_i^{\mathcal{M}'} v'$ and $v' \notin V^{\mathcal{M}'}(p)$ but then by Back there is a state $v \in Dom(\mathcal{M})$ such that $f(v) = v'$ and $w <_i^{\mathcal{M}} v$. But by Atomic Harmony, $v \notin V^{\mathcal{M}}(p)$ and thus *PrefL*11 is not satisfied at $\mathcal{M}, w$, concluding our proof. QED

Proposition 7.17. *On S/TPO-N-LTS, ST*1 *is invariant under* BM.

Proof. One direction follows directly from R-homomorphism. For the other direction, assume that *PrefL*11 is not satisfied at $\mathcal{M}', f(w)$. It follows that there is a state $v' \in Dom(\mathcal{M}')$ such that $f(w) \overset{\{i\}}{\rightarrow}' v'$ and $f(w) <_i^{\mathcal{M}'} v'$ (a). But then by Back there is a state $v \in Dom(\mathcal{M})$ such that $f(v) = v'$ and $w \overset{\{i\}}{\rightarrow} v$ (b). We are in one of two cases. Case 1: $v \not\leq_i^{\mathcal{M}} w$ (c). But then by totality we have $w \leq_i^{\mathcal{M}} v$ (d). But it follows from (b), (c), (d) that *PrefL*11 is not satisfied at $\mathcal{M}, w$. Now assume for a contradiction that we are in the other case. Case 2: $v \leq_i^{\mathcal{M}} w$ (e). But then by R-homomorphism, we have $f(v) \leq_i^{\mathcal{M}'} f(w)$ contradicting the assumption that (a) and concluding our proof. QED

Comments. The failures of invariance under GSM are still present, reflecting the fact that we do not have converse relations. By contrast, *PrefL*11 and *PrefL*2 are now invariant under bisimulation and a simple boolean modal logic with intersection seems to have the right expressive power to talk about efficiency and stability notions, since all of them are now invariant under ∩-Bisimulations. We

now check if the properties define classes of frames that are closed under the different operations introduced. The results can be read off the tables as in the previous section.

Closure Results for class of frames defined by global properties

	BMI GSF DU	refl.GSF BisSysIm		BMI	GSF DU	refl.GSF BisSysIm		BMI GSF DU	refl.GSF BisSysIm
*PowG*1	Y Y Y	Y Y	*PrefG*4	Y	Y N	N Y	*PPG*1	Y Y Y	Y Y
*PowG*2	Y Y Y	Y Y	*PrefG*5	N	Y Y	Y Y	*PPG*2	Y Y N	N Y
*PowG*3	Y Y Y	Y Y	*PrefG*6	N	Y N	N Y	*PPG*3	N N N	Y ?
*PowG*4	Y Y Y	Y Y	*PrefG*7	N	Y N	N Y	*PPG*4	N Y Y	Y Y
*PowG*5	Y Y Y	Y Y	*PrefG*8	Y	Y Y	Y Y	*PPG*5	Y Y Y	Y Y
*PrefG*1	Y Y Y	Y Y	*PrefG*9	N	Y Y	Y Y	*PPG*6	Y Y Y	Y Y
*PrefG*2	Y Y Y	Y Y	*PrefG*10	Y	Y Y	Y Y	*PPG*7	Y N Y	Y Y
*PrefG*3	Y Y N	N Y	*PrefG*11	N (18)	Y Y	Y Y			

Proposition 7.18. *Validity of PrefG*11 *is not preserved under* BMI.

Proof. Consider the frames $\mathcal{F} = \langle \{x, v, w\}, \{1, 2\}, \{ \overset{\mathtt{C}}{\rightarrow} \mid \mathtt{C} \subseteq \{1, 2\}\}, \{\leq_1, \leq_2\}\rangle$, with $w \overset{1}{\rightarrow} v, w \overset{2}{\rightarrow} v, \overset{\{1,2\}}{\rightarrow} = \emptyset, w \leq_1 v, w \leq_2 x$ and $\mathcal{F}' = \langle \{s, t\}, \{1, 2\}, \{ \overset{\mathtt{C}}{\rightarrow} \mid \mathtt{C} \subseteq \{1, 2\}\}, \{\leq'_1, \leq'_2\}\rangle$, with $s \overset{1}{\rightarrow} t, s \overset{2}{\rightarrow} t, \overset{\{1,2\}}{\rightarrow} = \emptyset, s \leq_1 t, s \leq_2 t$. Then $f : W \rightarrow W', f(w) = s, f(v) = f(x) = t$ is a surjective BM. However, $\mathcal{F} \Vdash$ *PrefG*11 and $\mathcal{F}' \nVdash$ *PrefG*11 because $s \leq_1 t, s \leq_2 t$ and it is not the case that $s \overset{\{1,2\}}{\rightarrow} t$. QED

At the frame level, ML is a fragment of Monadic Second Order Logic. That it does better at this level is thus not only an artifact of the chosen notions.

7.6 Modal definability

The previous model-theoretic results give us information about possible definability results. However, let us be more constructive and give formulas that indeed do the job: be it for local-satisfaction or frame-definability aims. We indicate the least expressive language we found still being able to express the property under consideration, the tightness of these results being given by our earlier invariance results. Another useful criterion is that of the computational complexity of the logic, i.e. of its satisfiability problem (SAT) and on its model checking problem (MC). We bridge our expressivity and complexity results as follows: for each local (resp. global) notion, we look for the least expressive logic that is still able to express it locally (resp. define the class of frames corresponding to it) and take the complexity of this logic as an *upper bound*. More precisely we take the upper bounds on its SAT problem and of (the combined complexity of) its model-checking. We indicate these upper bounds and references to the papers in

which these complexity results were proven. The reader who is not familiar with P, PSPACE and EXPTIME is referred to [129]. Π^0_1-complete problems [125] are undecidable but co-recursively enumerable (e.g. $\mathbb{N} \times \mathbb{N}$ tiling [97]). Let us now start with our definability results.

7.6.1 Defining local notions

The following table summarizes our definability results (for local notions) and their corollaries with respect to upper bounds of satisfiability and of model-checking.

	Axiom	Best Language	SAT	MC
*PowL*1	$\langle \mathtt{C} \rangle p$	$\mathcal{L}(\mathtt{N})$	PSPACE[113]	P[74]
*PowL*2	$\bigwedge_{\mathtt{C} \not\ni i} [\mathtt{C}] \neg p$	$\mathcal{L}(\mathtt{N})$	PSPACE[93]	P[74]
*PowL*3	$\downarrow x.[\mathtt{D}] \downarrow y.@_x \langle \mathtt{C} \rangle y$ (19)	$\mathcal{L}(\mathtt{N}, \downarrow, @, x)$	Π^0_1[55]	PSPACE [76]
*PrefL*1	$\langle \leq_i \rangle p$	$\mathcal{L}(\mathtt{N})$	PSPACE[113]	P[74]
*PrefL*2	$\downarrow x.\langle \leq_i \rangle (p \wedge [\leq_i] \neg x)$	$\mathcal{L}(\mathtt{N}, \downarrow, x)$	EXPTIME[56]	PSPACE[76]
*PrefL*3	$\downarrow x.\langle \bigcap_{i \in N} \leq_i \rangle (\bigvee_{j \in \mathtt{N}} [\leq_j] \neg x)$	$\mathcal{L}(\mathtt{N}, \downarrow, \cap, x)$	Π^0_1[55]	PSPACE
*PrefL*4	$\langle \leq_i{}^{-1} \rangle p$	$\mathcal{L}(\mathtt{N}, \downarrow, @, x)$	PSPACE	PSPACE [76]
*PrefL*5	$\downarrow x.\langle \leq_i{}^{-1} \rangle (p \wedge [\leq_i{}^{-1}] \neg x)$	$\mathcal{L}(\mathtt{N}, \downarrow, {}^{-1}, x)$	Π^0_1 [91]	PSPACE
*PrefL*6	$[(\leq_i \cap \overline{\leq_j}) \cup (\leq_j \cap \overline{\leq_i})] \bot$	$\mathcal{L}(\mathtt{N}, -, \cap)$	EXPTIME[120]	P[114]
*PrefL*7	$\langle \leq_i \cap (\bigcap_{j \in \mathtt{N} - \{i\}} \overline{\leq_j}) \rangle \top$	$\mathcal{L}(\mathtt{N}, -, \cap)$	EXPTIME[120]	P[114]
*PrefL*8	$[\![\leq_i]\!] p$	$\mathcal{L}(\mathtt{N}, [\![\ ]\!])$	EXPTIME[120]	P[114]
*PrefL*9	$\downarrow x.\mathtt{A} \downarrow y.(\neg \langle \leq_i \rangle x \wedge @_x \langle \leq_i \rangle y)$	$\mathcal{L}(\downarrow, @, x, \mathtt{E})$	Π^0_1 [91]	PSPACE[75]
*PrefL*10	$[\leq_i] p$	$\mathcal{L}(\mathtt{N})$	PSPACE[93]	P[74]
*PrefL*11	$\downarrow x.[\leq_i]([\leq_i] \neg x \Rightarrow p)$ (20)	$\mathcal{L}(\downarrow, x)$	(22)EXPTIME[56]	PSPACE[76]
*PPL*1	$\langle \mathtt{C} \cap \leq_i \rangle \top$	$\mathcal{L}(\mathtt{N}, \cap)$	PSPACE [68]	P[114]
*PPL*2	$\langle \mathtt{C} \cap (\bigcap_{i \in \mathtt{D}} \leq_i) \rangle \top$	$\mathcal{L}(\mathtt{N}, \cap)$	PSPACE[68]	P[114]
*PPL*3	$\langle (\bigcap_{i \in \mathtt{N}} \leq_i) \cap (\bigcup_{\mathtt{C} \subseteq \mathtt{N}} \overset{\mathtt{C}}{\rightarrow}) \rangle \top$	$\mathcal{L}(\mathtt{N}, -, \cap)$	EXPTIME[120]	P[114]
*PPL*4	$[\overline{\mathtt{C}} \cap \leq_i] \bot$	$\mathcal{L}(\mathtt{N}, -, \cap)$	EXPTIME[120]	P[114]
*PPL*5	$\downarrow x.[\overline{\mathtt{C}} \cap \leq_i] \langle \leq_i \rangle x$	$\mathcal{L}(\mathtt{N} \downarrow, , -, \cap, x)$	Π^0_1 [91]	PSPACE
*PPL*6	$\bigvee_{\mathtt{C} \subseteq \mathtt{N}} [\mathtt{C} \cap \overline{\leq_i}] \bot$	$\mathcal{L}(\mathtt{N}, -, \cap)$	EXPTIME[120]	P[114]
*PPL*7	$\downarrow x.[\mathtt{C}] \downarrow y.(\neg \langle \leq_i \rangle x \wedge @_x \langle \leq_i \rangle y)$	$\mathcal{L}(\mathtt{N}, \downarrow, @, x)$	Π^0_1[56]	PSPACE[76]
*EF*1	$\downarrow x.[\bigcap_{i \in \mathtt{N}} \leq_i] \bigvee_{i \in \mathtt{N}} \langle \leq_i \rangle x$	$\mathcal{L}(\mathtt{N}, \downarrow, \cap)$	Π^0_1[55, 56]	PSPACE
*EF*2	$\neg \downarrow x.\langle \bigcap_{i \in N} \leq_i \rangle (\bigvee_{j \in \mathtt{N}} [\leq_j] \neg x)$	$\mathcal{L}(\mathtt{N}, \downarrow, \cap)$	Π^0_1[55, 56]	PSPACE
*EF*3	$[\bigcap_{i \in \mathtt{N}} \leq_i] \bot$	$\mathcal{L}(\mathtt{N}, \cap)$	PSPACE [68]	P[114]
*ST*1	$\bigwedge_{i \in \mathtt{N}} \downarrow x.[i \cap \leq_i] \langle \leq_i \rangle x$	$\mathcal{L}(\mathtt{N}, \downarrow, \cap)$	Π^0_1[55]	PSPACE
*ST*2	$\bigwedge_{i \in \mathtt{N}} [i \cap \leq_i] \bot$	$\mathcal{L}(\mathtt{N}, \cap)$	PSPACE [68]	P[114]
*ST*3	$\bigwedge_{\mathtt{C} \subseteq \mathtt{N}} \downarrow x.[\mathtt{C} \cap (\bigcap_{i \in \mathtt{C}} \leq_i)] \bigvee_{j \in \mathtt{C}} \langle \leq_j \rangle x$	$\mathcal{L}(\mathtt{N}, \downarrow, \cap)$	Π^0_1[56]	PSPACE
*ST*4	$\bigwedge_{\mathtt{C} \subseteq \mathtt{N}} [\mathtt{C} \cap (\bigcap_{i \in \mathtt{C}} \leq_i)] \bot$	$\mathcal{L}(\mathtt{N}, \cap)$	PSPACE [68]	P[114]

Representative definability results.

Proposition 7.19. *PowL3 is true of $\mathcal{M}, w$ iff $\mathcal{M}, w, g \Vdash \downarrow x.[\mathtt{D}] \downarrow y.@_x \langle \mathtt{C} \rangle y$.*

Proof. From right to left: Assume that $\mathcal{M}, w, g \Vdash \downarrow x.[\mathtt{D}] \downarrow y.@_x \langle \mathtt{C} \rangle y$. Then we have $\mathcal{M}, w, g[x := w], \Vdash [\mathtt{D}] \downarrow y.@_x \langle \mathtt{C} \rangle y$. But now assume there is a state v that coalition

$\mathtt{D}$ can force from w. By definition, $w \xrightarrow{\mathtt{D}} v$ (1). But by (1) and semantics of $[\mathtt{D}]$ then we have $\mathcal{M}, v, g[x := w], \Vdash \downarrow y.@_x\langle \mathtt{C}\rangle y$ (2). (2) and semantics of $\downarrow$ gives us $\mathcal{M}, v, g[x := w, y := v] \Vdash @_x\langle \mathtt{C}\rangle y$ (3). From (3) and semantics of $@_x$ and the fact that $g(x) = w$ we have $\mathcal{M}, w, g[x := w, y := v] \Vdash \langle \mathtt{C}\rangle y$ (4). But by semantics of $\langle \mathtt{C}\rangle$ and the fact that $g(y) = v$, (4) really means that $w \xrightarrow{\mathtt{C}} v$ (5). Since the v was arbitrary, it follows from (5) that at w for any state v, if $\mathtt{D}$ can achieve it, then $\mathtt{C}$ can do so, too. But this precisely means that *PowL3* is true of $\mathcal{M}, w$. QED

Proposition 7.20. *PrefL*11 *is true of* $\mathcal{M}, w$ *iff* $\downarrow x.[\leq_i]([\leq_i]\neg x \Rightarrow p)$.

Proof. From right to left: Assume that $\mathcal{M}, w, g \Vdash \downarrow x.[\leq_i]([\leq_i]\neg x \Rightarrow p)$. Then we have $\mathcal{M}, w, g[x := w], \Vdash [\leq_i]([\leq_i]\neg x \Rightarrow p)$ (1). Take an arbitrary state v such that $w <_i v$ (2). We will prove that v is a p-state. It follows from (2) that $w \leq_i v$ (3). From (1), (3) and semantics of $[\leq_i]$ it follows that $\mathcal{M}, v, g[x := w], \Vdash [\leq_i]\neg x \Rightarrow p$ (4). But by (2) we have $v \not\leq_i w$. It follows by semantics of x and $[\leq_i]$ that $\mathcal{M}, v, g[x := w], \Vdash [\leq_i]\neg x$. By (4) and (5) it follows that $v \in V(p)$. QED

Theorem 7.21 (ten Cate and Franceschet [56])**.** *The satisfiability problem for formulas in the modal language* $\mathcal{L}(\mathtt{N}, \downarrow, @, x) - \Box\downarrow\Box$ *with bounded width is* EXPTIME*-complete.*

Proposition 7.22. *PrefL*11 *is expressible in an extended modal language with a satisfiability problem in* EXPTIME.

Proof. By the previous proposition, we have *PrefL*11 is defined by $\downarrow x.[\leq_i]([\leq_i]\neg x \Rightarrow p)$. But $\downarrow x.[\leq_i]([\leq_i]\neg x \Rightarrow p)$ does not contain the $\Box\downarrow\Box$ scheme. Thus, *PrefL*11 is defined by a formula in $\mathcal{ML}(\mathtt{N}, \downarrow, @, x) - \Box\downarrow\Box$ (1). But by Theorem 7.21 the satisfiability problem of $\mathcal{ML}(\mathtt{N}, \downarrow, @, x) - \Box\downarrow\Box$ is in EXPTIME. QED

On the model-theoretic level we observed that our notions were generally not invariant under taking bounded morphic images. This is reflected in the fact that, on the syntactic level, most of the notions use intersection: be it to require a state to be better for all agents or to check whether a better state of the system can be achieved. But in general a boolean modal logic with complement and intersection can express a lot of relevant notions, suggesting that social-choice and game-theoretical reasoning in a modal logic setting about models in which the representation of coalitional powers is greatly simplified need not require logics with a SAT problem worse than EXPTIME. Finally it is interesting to note that strong notions of stability (strong Nash stability) or of efficiency (strong Pareto-effiency) are easier to express than their weak counterparts (Nash stability and Pareto-efficiency) for which a decidable modal logic needs to be looked for outside the space of logics we have been considering, using a more refined scale of candidate modal logics. We now turn to the global notions.

7.6.2 Defining global notions

We now present our findings for global notions and similarly their corollaries in terms of upper bounds on the complexity of a modal logic that can define them.

	Axiom	Best Language	SAT	MC
*PowG*1	$\bigwedge_{\mathtt{C}\subseteq\mathtt{N}}((\langle\mathtt{C}\rangle\varphi \Rightarrow [\mathtt{C}]\varphi) \wedge \langle\mathtt{C}\rangle\top)$	$\mathcal{L}(\mathtt{N})$	PSPACE[113]	P[74]
*PowG*2	$\bigwedge_{\mathtt{C}:\lvert\mathtt{C}\rvert<\lvert\mathtt{N}\rvert/2}[\mathtt{C}]\bot$	$\mathcal{L}(\mathtt{N})$	PSPACE[93]	P[74]
*PowG*3	$\bigwedge_{\mathtt{C}\subseteq\mathtt{N}}(\langle\mathtt{C}\rangle\varphi \Rightarrow [\mathtt{C}]\varphi)$	$\mathcal{L}(\mathtt{N})$	PSPACE[113]	P[74]
*PowG*4	$\bigwedge_{\mathtt{C}\subseteq\mathtt{N}}\bigwedge_{\mathtt{D}\supseteq\mathtt{C}}(\langle\mathtt{C}\rangle\varphi \Rightarrow \langle\mathtt{D}\rangle\varphi)$	$\mathcal{L}(\mathtt{N})$	PSPACE[93]	P[74]
*PowG*5	$\langle\mathtt{C}\rangle\top \Rightarrow \bigwedge_{\mathtt{D}:\mathtt{C}\cap\mathtt{D}=\emptyset}(\langle\mathtt{D}\rangle\varphi \Rightarrow \langle\mathtt{C}\rangle\varphi)$	$\mathcal{L}(\mathtt{N})$	PSPACE[113]	P[74]
*PrefG*1	$\varphi \Rightarrow \langle\leq_i\rangle\varphi$	$\mathcal{L}(\mathtt{N})$	PSPACE[93]	P[74]
*PrefG*2	$\langle\leq_i\rangle\langle\leq_i\rangle\varphi \Rightarrow \langle\leq_i\rangle\varphi$	$\mathcal{L}(\mathtt{N})$	PSPACE[113]	P[74]
*PrefG*3	$(p \wedge \mathtt{E}q) \Rightarrow (\mathtt{E}(p \wedge \langle\leq_i\rangle q) \vee \mathtt{E}(q \wedge \langle\leq_i\rangle p))$	$\mathcal{L}(\mathtt{N},\mathtt{E})$	EXPTIME[144]	P[114]
*PrefG*4	Conjunction of the 3 previous axioms	$\mathcal{L}(\mathtt{N},\mathtt{E})$	EXPTIME[101]	P[114]
*PrefG*5	see below	$\mathcal{L}(\mathtt{N})$	PSPACE[93]	P[74]
*PrefG*6	$\bigwedge_{i\in\mathtt{N}}(@_j\langle\leq_i\rangle k \vee @_k j \vee @_k\langle\leq_i\rangle j)$	$\mathcal{L}(\mathtt{N},@,i)$	PSPACE[8]	P[76]
*PrefG*7	*PrefG*5 $\wedge$ *PrefG*6 $\wedge$ $(\bigwedge_{i\in\mathtt{N}}(j \Rightarrow \neg\langle\leq_i\rangle j))$	$\mathcal{L}(\mathtt{N},@,i)$	PSPACE[8]	P[76]
*PrefG*8	$\bigwedge_{i\in\mathtt{N}}\ ((\langle\leq_i\rangle\varphi \Rightarrow [\leq_i]\varphi) \wedge \langle\leq_i\rangle\top)$	$\mathcal{L}(\mathtt{N})$	PSPACE[93]	P[74]
*PrefG*9	$\langle\mathtt{C}\rangle j \leftrightarrow \bigwedge_{i\in\mathtt{C}}\langle\leq_i\rangle j$	$\mathcal{L}(\mathtt{N},i)$	PSPACE[8]	P[76]
*PrefG*10	$\langle\mathtt{C}\rangle p \leftrightarrow \bigvee_{i\in\mathtt{C}}\langle\leq_i\rangle p$	$\mathcal{L}(\mathtt{N})$	PSPACE[113]	P[74]
*PrefG*11	$\langle\mathtt{C}\rangle j \leftrightarrow \bigvee_{\mathtt{D}\subseteq\mathtt{C}\&\lvert\mathtt{D}\rvert>\frac{\lvert\mathtt{C}\rvert}{2}}(\bigwedge_{i\in\mathtt{D}}\langle\leq_i\rangle j)$	$\mathcal{L}(\mathtt{N},i)$	PSPACE[8]	P[76]
*PPG*1	$\langle\mathtt{C}\rangle\varphi \Rightarrow \bigwedge_{i\in\mathtt{N}}\langle\leq_i\rangle\varphi$	$\mathcal{L}(\mathtt{N})$	PSPACE[93]	P[74]
*PPG*2	$\bigvee_{i\in\mathtt{N}}\mathtt{A}\bigwedge_{\mathtt{C}\subseteq\mathtt{N}}(\langle\mathtt{C}\rangle\varphi \Rightarrow \langle\leq_i\rangle\varphi)$	$\mathcal{L}(\mathtt{N},\mathtt{E})$	EXPTIME[144]	P[114]
*PPG*3	$\bigwedge_{i\in\mathtt{N}}\bigvee_{\mathtt{C}\subseteq\mathtt{N}}(\overline{\leq_i\cup\leq_i})(\overline{\leq_i}\cap\mathtt{C})\top$	$\mathcal{L}(\text{-},\cap,\cup)$	EXPTIME[120]	P[114]
*PPG*4	see below	$\mathcal{L}(\mathtt{N},i)$	PSPACE[8]	P[76]
*PPG*5	$\bigwedge_{\mathtt{C}\not\supseteq\{i\}}[\overset{\mathtt{C}}{\rightarrow}]\bot$	$\mathcal{L}(\mathtt{N})$	PSPACE[113]	P[74]
*PPG*6	$\langle\mathtt{C}\rangle\varphi \Rightarrow \bigvee_{\mathtt{D}\subset\mathtt{C}}\langle\mathtt{D}\rangle\varphi$	$\mathcal{L}(\mathtt{N})$	PSPACE[93]	P[74]
*PPG*7	$\bigwedge_{i\in\mathtt{N}}\mathtt{E}\bigwedge_{\mathtt{C}\not\supseteq\{i\}}[\mathtt{C}]\bot$	$\mathcal{L}(\mathtt{N},\mathtt{E})$	EXPTIME[8]	P[76]

$$\bigwedge_{i\in\mathtt{N}}(j \wedge \langle\leq_i\rangle(k \wedge \neg\langle\leq_i\rangle j \wedge \langle\leq_i\rangle(l \wedge \neg\langle\leq_i\rangle k))) \Rightarrow j \wedge \langle\leq_i\rangle(l \wedge \neg\langle\leq_i\rangle j) \qquad (AxPrefG5)$$

$$[p \wedge \langle i\rangle q \wedge \langle\leq_i\rangle(q \wedge \langle\leq_i\rangle\neg p)] \Rightarrow \bigwedge_{i\in\mathtt{C}\subseteq\mathtt{N}}[(\langle\mathtt{C}\rangle r \wedge \bigwedge_{\mathtt{D}\subseteq\mathtt{C}\setminus i}\neg\langle\mathtt{D}\rangle r) \Rightarrow \langle\leq_i\rangle(r \wedge \neg\langle\leq_i\rangle p)] \ (PPG4)$$

From our earlier closure results for these notions there is no suprise in the fact that most of these notions are definable in the basic modal languages, corresponding to the earlier mentioned fact that on the level of frames modal logic is really a fragment of MSO. More generally global notions will generally not be responsible for the complexity cost when designing a modal language to reason about coalitional power and preferences. This remark is to be balanced by the fact that we generally use very big conjunctions or disjunctions that might well be exponential if we take the number of agents as a parameter for the complexity results.

The conclusion will give a bigger picture. For now let us indicate, based on our current work, lines that seem worth exploring:

- Since dealing with real coalitional powers is probably more natural using neighborhood semantics, it will be useful to do the same work for modal logics of the **CL**-type or of the type of one of its normal simulations [53].

- More generally the preceding invariance and definability results (as well as their corollaries in terms of complexity) depend on the choice of models. [63] investigates how the expressive power and complexity requirements are affected by the choice of models.

- It would be interesting to obtain similar invariance results and upper bounds on the complexity of the logics needed to encode *concrete arguments* from SCT and (cooperative) GT, thus addressing the complexity of *actual reasoning* about cooperative situations.

- From our definability results we could obtain upper bounds (on SAT) and on the *combined complexity* of model checking of logics able to express certain notions from SCT and GT. The converse road would be to use complexity results from computational social choice and algorithmic game theory to obtain lower bounds on its *data complexity*.[3] As an example: a way to go could be to take a hardness result for the problem of determining whether a profile of strategies is a pure Nash equilibrium of a given game (with respect to some reasonable and qualitative encoding of games) as a lower bound on the data complexity of model-checking of a logic that can express this notion.

- In order to obtain a complete picture of the complexity of reasoning about cooperation, we need a procedure to assess the LB of the complexity of modal logics that can express some notion.

- Our definability results made use of very big conjunctions and disjunctions. It would be interesting to check how the length of these formulas is related to a more reasonable input such as the number of agents. (Taking conjunctions/disjunctions over all coalitions, they will be exponentially related.)

- We could also consider the complexity effects of using more succinct languages that have more modalities, e.g. a modality $\langle Most \leq \rangle \varphi$, read: "there is a φ-state that a majority of agents finds at least as good as the current one" (cf. e.g. [2]).

[3] When measuring combined complexity both the formula and the model are part of the input, while when measuring data complexity, the formula is fixed and the model is the input (see [151]).

7.7 Conclusion

We investigated how reasoning about coalitional powers and preferences in a modal logic setting is demanding in expressive power, by giving invariance and definability results for notions inspired from game theory and social choice theory with respect to models in which the representation of coalitional power is greatly simplified.

Major sources. The point of departure of this chapter was the recent development of different modal logical systems involving both preferences and coalitional power in a number of works such as Kurzen [112], Ågotnes et al. [1]. The particular models we chose can be seen as a generalization of Segerberg [143] for coalitions instead of single agents (complemented with preferences). The notions of a Pareto-efficient and of a Nash-stable state were considered independently in [140]. Finally the tightness of our definability results, given our invariance results, is backed up by characterization and invariance results for (extended) modal languages; we refer to Appendix B for references. The papers proving the complexity results for different modal languages from which we obtain our corollaries are too numerous to list here, the tables have pointers to the relevant papers.

Our main results. This chapter identified natural notions for reasoning about cooperation: local notions giving properties of a state of a system and global notions defining a class of frames. We provided satisfiability (resp. validity) invariance results for these notions for a large class of operations and relations between models (resp. frames). We also gave explicit definability results and observed, on the one hand, that defining frames for cooperation logics is not too demanding in terms of expressive power, as most of the notions considered are definable in the basic modal language. On the other hand, our results show that local notions call for modal logics for which satisfaction is not invariant under bounded morphisms. However, as long as we avoid converse modalities, interesting reasoning about cooperation can be done within GSM-invariant modal languages. Though this fact does not directly lead to a nice upper bound on the complexity of the logic's SAT (nor to its decidability), our definability results show that most of the considered notions can (individually) be expressed in MLs in EXPTIME. Moreover, for several notions we only found logics with undecidable SAT that could express them. All these notions involve the idea of a "strict" improvement (e.g. Nash-stable, Pareto-efficient). By contrast, strong notions of stability and efficiency (EF3, ST2, ST4) are all expressible in logics with SAT in PSPACE. Thus, we could say that "expressing strictness" and therefore "weak" notions are dangerous, while "strong" notions (looking only at the weak preference relation) are safe.

The last step. This chapter has turned the focus from belief change to preferences and coalitional power, extending our logical analysis based on modal logics to a complementary dimension of intelligent interaction. The next chapter concludes this dissertation.

Chapter 8

Conclusion: reasoning about reasoning

Modeling rational agents' reasoning in interactive contexts and identifying its logic is the general analytic project to which this dissertation contributes. The borders of this project run through economics, computer science and philosophy. It includes several theoretical lines that we have connected. Interactive epistemology is the study of interactive reasoning: how agents entertain beliefs and reason about the beliefs of other agents. Formal learning theory is the study of the conditions under which agents can reach stable beliefs or identify a correct hypothesis from a stream of data. Epistemic game theory is a theory of how rational agents would make decisions based on their beliefs in strategic interactive situations. In all these systems, beliefs, interactive beliefs, and their evolution as informational processes unfold are at stake.

This dissertation has connected these themes by developing one single logical framework. For this purpose, we operated at the interface of two major logics of belief change: the temporal approach and the dynamic approach. Concretely, we connected and merged the two families of logics, first at a structural semantic level and then at a syntactic one. Subsequently, we applied the resulting system to analyze what happens to agents' beliefs over time when agents communicate, learn, interact, and reason interactively, inductively, or strategically.

Chapter 2 identified the main structural properties of belief revising agents over time, and Chapter 3 then formulated its main logical proof principles. This chiefly took the form of semantic representation theorems, plus a completeness theorem for changing beliefs in a temporal logic that admits protocols. Chapter 4 identified common belief of posteriors in suitable structures as a key sufficient condition for agents to reach agreement, and iterated announcement of beliefs as a major way of reaching it. We also determined the right family of static and dynamic logics to reason about agreement, and found agreement results, invariance results, and concrete syntactic proofs of agreement results. Chapter 5 investigated the logical principles behind inductive learning and in particular behind the key notion of finite identifiability. This took the form of a reduction

of the problem of identifiability to a problem of model-checking for an epistemic temporal logic, plus further representation results. Chapter 6 took the dynamic-temporal logical viewpoint to the building blocks of strategic reasoning: solution algorithms, rationality, equilibrium, and expectations, discussing the importance of belief change for the epistemic foundations of game theory. We gave many concrete scenarios sketching a bigger picture. Chapter 7 completed the whole approach with two further key aspects of agency: preferences, and coalitional powers. We explored the logical expressive power demanded by notions imported in this area from social choice theory and cooperative and non-cooperative game theory, in terms of modal invariance and definability.

Some major open problems

This dissertation has made a number of connections, but in doing so, it raises an even larger number of new issues. From where we stand now, what are the most important tasks that lie ahead?

Question 8.1. *Can agreement be reached via soft updates?*

Public announcements of beliefs really represent a type of disagreement-solving scenario in which agents take belief announcement as hard information. In the preceding chapters we have extensively discussed soft dynamics such as lexicographic upgrade. What if agents instead of eliminating incompatible states, simply re-arrange their plausibility ordering: will they reach agreement in the limit? Does one need to make stronger assumptions about their prior beliefs? Will in general more radical or more conservative types of updates bring faster convergence?

Question 8.2. *What are the logical principles of function learning?*

Question 8.3. *What are the logical principles of identification in the limit?*

The connection between modal logics of belief change and learning theory is just emerging. We focused on a very specific type of learning: with languages viewed as sets of natural numbers, and a very specific notion of convergence: finite identifiability. The next step would be an analysis of function learning and less demanding notions of convergence such as identification in the limit. The full generality of soft updates based on plausibility event models and priority update might be useful for the study of identification in the limit. Function learning sounds somewhat more challenging for a stepwise approach to belief change, since the environments are no longer closed under permutation and might call for history-dependent learning strategies.

Question 8.4. *What is the complete logic of expectations?*

Question 8.5. *What is the complexity of the logic of expectations?*

We gave an outline for a logic of expectations in branching time. Technically, its semantics has features of other logics of agency, but does not seem to match them precisely. It would be of interest to obtain an axiomatic completeness result, or a proof of higher complexity. Axiomatic completeness results for CTL^* or STIT might be relevant here.

Question 8.6. *What are the logical principles of dynamic agreement results based on public announcement?*

We stated static agreement results and gave a syntactic proof matching one of them. We also stated dynamic results, with conditions under which agents will end up agreeing. The syntactic counterpart for these dynamic results might live in the region of inflationary modal fixed-point logics. But where exactly is a major open question.

Question 8.7. *What are reasonable notions of identifiability under imperfect observation?*

Question 8.8. *What happens when learners can communicate?*

Formal learning theory usually focus on learning situations in which learners try to converge to some hypothesis from unambiguous data. What happens if instead they are replaced by signals whose interpretation might be unclear or that might be noisy? What would be reasonable concepts of identifiability in such situations? We suggested that communication between learners plays a role in these contexts. Can this role be analyzed systematically?

Question 8.9. *What are the effects of the choice of protocol on identifiability of sets?*

Question 8.10. *What are the effects of the choice of protocol on dynamic agreement results?*

Question 8.11. *Can argumentative and dialogical scenarios be analyzed from the perspective of 'dynamic-temporal' logics?*

Restricting attention to particular classes of protocols might lead to scenarios in which learning is impossible or on the contrary trivial. The same is true for reaching agreements. In between are many interesting non-equivalent protocols. Can the effect of varying the protocols be studied in a systematic manner? Very concrete protocols are those considered by some argumentative and dialogical games. Is it possible to account systematically for some of these games from the perspective of dynamic-temporal logic of belief change?

Question 8.12. *How can one define lower bounds for the complexity of* classes *of logics of cooperation?*

Question 8.13. *What are the lower bounds of the data complexity of model checking notions from social choice and game theory when the size of the input is taken to be that of the* model*? — of the number of possible allocations? — of the number of profiles of pure strategies?*

From definability results we obtained upper bounds on the combined complexity of model checking for modal logics of cooperation. Our invariance results showed that these results are tight to some extent. It would be interesting to also find lower bounds on the *data complexity* of model checking these notions. Results in the computational social choice literature or in algorithmic game theory should be relevant here. But they usually take the number of agents or resources as input size, while a logical approach would take size of the model, generally exponentially bigger, calling for logarithmic hardness results.

More specific open problems have been stated at many places in this dissertation. We mention just a few:

- extend the completeness proof of Chapter 3 to interactive doxastic notions such as common belief,
- find a logic in which static agreement results can be derived syntactically, that would be finitely axiomatizable,
- extend the automatic Sahlqvist correspondence technique to doxastic and dynamic formulas, and similarly, for other correspondence algorithms such as structure-seeking dialogues.

A final thought

The outline of a logical theory of rational agency and intelligent interaction has been sharpening continuously in recent research. This dissertation has contributed its share, focusing on informational processes and logics of belief change over time. In our view, this logical theory is not meant to be a lonely helium balloon.

Loneliness. A logical theory of rational interaction is not meant to stand by itself, since it can only work properly in interaction with other fields that analyze human reasoning and human interaction. This dissertation has built connections in that spirit.

Helium balloon. It is not meant to be a purely analytical theory dealing with how theoretical agents *would* reason and interact, but also to explain how real agents with their cognitive limitations *do* reason and interact. But if so, it cannot ignore empirical data about real people. Giving a logical account of the cognitive limitations of real agents might sound like a recurring open problem but we do think that moving towards reality is one of the most challenging and exciting steps.

Appendix A

Some basics of interactive epistemology

We introduce the framework considered by [13, 11, 12] and Osborne and Rubinstein [127, ch.5] to encode the beliefs and information (or knowledge) of the agents.

Definition A.1 (Information function)**.** *Given a set of states Ω let an information function be a function $\mathbf{I} : \Omega \to \wp(\Omega)$ such that*

1. *$\omega \in \mathbf{I}(\omega)$ for every $\omega \in \Omega$*
2. *$\mathbf{I}(\omega) = \mathbf{I}(\omega')$ whenever $w' \in \mathbf{I}(w)$*

An information function really induces a partition on Ω.

Definition A.2 (Aumann probabilistic model of knowledge and beliefs)**.** *Given a countable set of states Ω and a finite set of agents N, a probabilistic model for knowledge and belief is of the form $\langle \Omega, (\mathbf{I}_i), \Sigma, N, (\pi_i)_{i \in N} \rangle$ where for each $i \in N$, $\mathbf{I}_i : \Omega \to \wp(\Omega)$ is an* information function*, Σ is a σ-algebra over Ω such that for all $i \in N$ and $\omega \in \Omega$, $\mathbf{I}_i(\omega) \in \Sigma$, and π_i is a (prior) probability measure on Ω.*

Elements of Σ are called *events*. They can be thought of as the natural semantic counterpart of a non-doxastic, non-epistemic formula, e.g. the time at which a particular movie is played in a given theater. Intuitively $\mathbf{I}_i(\omega)$ encodes the information or knowledge of i at ω. Agent i knows that E at ω if and only if $E \subseteq \mathbf{I}_i(\omega)$. π_i gives the prior (probabilistic) beliefs of i. Finally the (probabilistic) beliefs of agent i at ω are obtained by conditioning on his information function. Formally the posterior probability that i assigns to an event $E \in \Sigma$ at a state ω is

$$\pi_i(E \mid \mathbf{I}_i(\omega)) = \frac{\pi_i(E \cap \mathbf{I}_i(\omega))}{\pi_i(\mathbf{I}_i(\omega))}$$

Before we state an important result about such probabilistic models of knowledge and beliefs we need to make a few remarks and introduce a new notion. It is easy to see that the agents' information (or knowledge) could equivalently be encoded

by an equivalence relation and doing so is actually standard within qualitative approaches this dissertation is in line with and as introduced in Subsection 1.3.2. In this context we introduced a notion of *common knowledge* which is different but equivalent on partitional (and relational) structures to the one Aumann [13] is using. We will state Aumann's definition for the case of two agents.

Let E be an event. Intuitively it is common knowledge between agent 1 and agent 2 that E iff 1 knows that E, 2 knows that E, 1 knows that 2 knows that E, 2 knows that 1 knows that E, 1 knows that 2 knows that 1 knows that E and so on. We use the equivalent definition of Osborne and Rubinstein [127], ch.5

Definition A.3 (Osborne and Rubinstein [127, ch.5])**.** *An event $F \in \Sigma$ is* self-evident *if for $i = 1, 2$ and for all $\omega \in F$, $\mathbf{I}_i(\omega) \subseteq F$.*

Definition A.4 (Aumann [13], Osborne and Rubinstein [127, ch.5])**.** *An event $E \in \Sigma$ is common knowledge between 1 and 2 at ω if there is a self-evident event F such that $\omega \in F \subseteq E$.*

We can now state an important result due to Aumann.

Theorem A.5 (Aumann [13])**.** *Suppose that Ω is finite and that agents 1 and 2 have the same prior belief (probability measure). If the posteriors that 1 and 2 assign to an event E are common knowledge between them at a state ω, then these posteriors must be equal.*

Appendix B

Some basics on modal definability and invariance

One can ask two different things about a modal language with respect to definability. The first one is whether it is able to distinguish between two given relational models with a distinguished state (pointed models). In this case invariance results can help us (Section B.1). The second one is whether the language is able to define a class of frames of interest. For this task one can draw on closure results (Section B.2). In general Blackburn et al. [39] and ten Cate [55] are very useful sources and the interested reader is referred to them for details.

B.1 Distinguishing pointed models

Given a modal language and two situations for which we have an adequate representation as pointed models, how can we determine if some formula of the language will be able to make the distinction? An equivalent question is whether there is a formula φ in this particular language such that $\mathcal{M}, w \Vdash \varphi$ iff $\mathcal{M}, w$ has some property of interest. Depending on the language we can draw on existing characterization results that establish invariance criterion for definability. We give two classical characterization results and refer to [39, 55] for details and additional results. And then we introduce the relevant operations and relations on pointed models that we will be using in this dissertation.

Background results. We indicate two classical characterization results. For details see [39, 55]. Let $\varphi(x)$ be a formula of the FO correspondence language with at most one free variable. [27] proved that $\varphi(x)$ is invariant under bisimulations iff $\varphi(x)$ is equivalent to the standard translation of a modal formula. While [9, 73] proved that $\varphi(x)$ is invariant under taking generated submodels iff $\varphi(x)$ is equivalent to the standard translation of a formula of $\mathcal{L}(\mathtt{N}, \downarrow, @, x)$.

We first introduce some relations between models. Let τ be a finite modal similarity type with only binary relations. Let $\mathcal{M} = \langle W, (R_k)_{k\in\tau}, V\rangle$ and $\mathcal{M}' =$

$\langle W', (R'_k)_{k\in\tau}, V'\rangle$ be models of similarity type τ.

Definition B.1 (Bisimulations)**.** *A bisimulation between $\mathcal{M}$ and $\mathcal{M}'$ is a non-empty binary relation $Z \subseteq W \times W'$ fulfilling the following conditions:*

`Atomic Harmony`	*For every $p \in$ PROP, wZw' implies $w \in V(p)$ iff $w' \in V'(p)$.*
`Forth`	*$\forall k \in \tau$, if wZw' & R_kwv then $\exists v' \in W'$ s.t. $R'_kw'v'$ & vZv'.*
`Back`	*$\forall k \in \tau$, if wZw' & $R'_kw'v'$ then $\exists v \in W$ s.t. R_kwv & vZv'.*

In a nutshell, $\cap$-bisimulations (resp. Cbisimulations) require that `Back` and `Forth` also hold for the intersection (resp. the converse) of the relations. $\mathcal{H}$-Bisimulations extend `Atomic Harmony` to nominals. Tbisimulations ($\mathcal{H}(@)$-bisimulations) are total[1] bisimulations (resp. total $\mathcal{H}$-bisimulations). $\mathcal{H}(\mathtt{E})$-bisimulations are $\mathcal{H}$-bisimulations matching states "with the same name". See [55] for details. We now define bounded morphisms, generated subframes and disjoint unions.

Definition B.2 (`BM`)**.** *$f : W \to W'$ is a bounded morphism from $\mathcal{M}$ to $\mathcal{M}'$ iff:*

`Atomic Harmony`	*For every $p \in$ PROP, $w \in V(p)$ iff $f(w) \in V'(p)$.*
`R-homomorphism`	*$\forall k \in \tau$, if R_kwv then $R'f(w)f(v)$.*
`Back`	*$\forall k \in \tau$, if $R'_kf(w)v'$ then $\exists v \in W$ s.t. $f(v) = v'$ and R_kwv.*

Definition B.3 (Generated submodel)**.** *We say that $\mathcal{M}'$ is a generated submodel (`GSM`) of $\mathcal{M}$ iff $W' \subseteq W$, $\forall k \in \tau$, $R'_k = R_k \cap (W' \times W')$, $\forall p \in$ PROP, $V'(p) = V(p) \cap (W' \times W')$ and if $w \in W'$ and R_kwv then $v \in W'$.*

In some cases we will be interested in the submodel generated by a particular subset A of the domain.

Definition B.4 (A-Generated submodel)**.** *Let us give the definition for the concrete case of* epistemic plausibility models *$\mathcal{M} = \langle W, (\leq_i)_{i\in N}, (\sim_i)_{i\in N}, V\rangle$ and $A \subseteq W$. The submodel of $\mathcal{M}$ generated by A (or A-generated submodel), that we write $\mathcal{M}^A$ is defined as follows: $\mathcal{M}^{\mathcal{A}} = \langle W^A, (\leq_i^A)_{i\in I}, (\sim_i^A)_{i\in I}, V^A\rangle$, where:*

- *$W^A = W \cap A$;*
- *For each $i \in N$, $\leq_i^A = \leq_i \cap (W^A \times W^A)$;*
- *For each $i \in N$, $\sim_i^A = \sim_i \cap (W^A \times W^A)$;*
- *For each $v \in W^A$, $v \in V^A(p)$ iff $v \in V(p)$.*

We write $Sub(\mathcal{M}) = \{\mathcal{M}'$ is the A-generated submodel of $\mathcal{M} \mid A \subseteq |\mathcal{M}|\}$ and $\mathcal{M}' \sqsubseteq \mathcal{M}$ whenever $\mathcal{M}' \in Sub(\mathcal{M})$.

[1] $Z \subseteq W \times W'$ is *total* iff $\forall w \in W\ \exists w' \in W'\ wZw'$ & $\forall w' \in W'\ \exists w \in W\ wZw'$.

Definition B.5 (Disjoint Unions)**.** *Let $(\mathcal{M}_j)_{j\in J}$ be a collection of models with disjoint domains. Define their disjoint union $\biguplus_j \mathcal{M}_j = \langle W, R, V\rangle$ as the union of their domains and relations, and define for each $p \in$ PROP, $V(p) := \bigcup_j V_j(p)$.*

Definition B.6 (Invariance)**.** *A property of pointed models $\Phi(X, y)$ is* invariant under λ-Bisimulations *iff whenever there exists a λ-bisimulation Z between $\mathcal{M}$ and $\mathcal{M}'$ such that $(w, w') \in Z$, then $\Phi(\mathcal{M}, w')$ holds iff $\Phi(\mathcal{M}', w')$ holds. Invariance for other operations is defined similarly.*

Similar results and tools can help us determine whether a modal language can define a given class of frames.

B.2 Defining classes of frames

First of all, we define what it means for a formula to be valid on a class of frames.

Definition B.7 (Validity on a class of frames)**.** *We say that a formula φ is valid on a class of frames $\mathbf{F}$ iff for any frame $\mathcal{F} \in \mathbf{F}$ and any model $\mathcal{M}$ based on $\mathcal{F}$, at all states w in $Dom(\mathcal{F})$, $\mathcal{M}, w \Vdash \varphi$. We write $\mathbf{F} \Vdash \varphi$.*

By defining a class of frames $\mathtt{F}$, we mean finding a formula φ such that for every frame $\mathcal{F}$ we have $\mathcal{F} \Vdash \varphi$ iff $\mathcal{F} \in \mathtt{F}$.

In the case of definability of classes of frames we will be interested in closure conditions. In fact we will mostly be interested in using the following result to prove that certain classes of frames are not definable in basic modal languages.

Background result. We indicate one classical characterization result on the level of frames. For details see [39, 55]. Goldblatt and Thomason [88] proved that a first-order definable class of frames is modally definable iff it is closed under taking `BMI`, `GSF`, disjoint unions and reflects ultrafilter extensions.

Let us now introduce the definitions of the relevant closure conditions. First, we consider bounded morphic images (`BMI`) of frames. `BM` on frames are obtained by dropping `Atomic Harmony` in Def. B.2. A class of frames is closed under `BMI` iff it is closed under *surjective* `BM`. Next, we consider closure under generated subframes (`GSF`) — the frame-analogue to `GSM` (cf. Def. B.3). We also check if properties *reflect* `GSF`. A property φ *reflects* `GSF` if whenever for every frame $\mathcal{F}$, it holds that every `GSF` of $\mathcal{F}$ has property φ, then so does $\mathcal{F}$. We also consider closure under taking disjoint unions (`DU`) of frames, which are defined in the obvious way. Moreover, we look at closure under images of bisimulation systems [55], which are families of partial isomorphisms.

Definition B.8 (Bisimulation System)**.** *A bisimulation system from a frame $\mathcal{F}$ to a frame $\mathcal{F}'$ is a function $\mathcal{Z} : \wp(W') \to \wp(W \times W')$ that assigns to each $Y \subseteq W'$ a total bisimulation $\mathcal{Z}(Y) \subseteq W \times W'$ such that for each $y \in Y$:*

1. *There is exactly one $w \in W$ such that $(w, y) \in \mathcal{Z}(Y)$.*
2. *If $(w, y), (w, w') \in \mathcal{Z}(Y)$, then $w' = y$.*

Appendix C

Additional proofs for Chapter 2

We prove a general representation theorem similar to the one in Section 2.7 and 1.8, for unified (or local) plausibility models (as introduced in Definition 1.17). Representing the local priority updaters (cf. Definition 2.24) in unified (or local) doxastic temporal terms. We recall Definition 2.24.

Definition. *The* local priority update *of a* unified *doxastic plausibility model* $\mathcal{M} = \langle W, (\trianglelefteq_i)_{i\in N}, V\rangle$ *and a* $\trianglelefteq$*-event model* $\epsilon = \langle E, (\trianglelefteq_i)_{i\in N}, \mathtt{pre}\rangle$ *is the unified plausibility model* $\mathcal{M}\otimes\epsilon = \langle W', (\trianglelefteq'_i)_{i\in N}, V'\rangle$ *constructed as follows:*

- $W' = \{(w,e) \in W \times E \mid \mathcal{M}, w \Vdash \mathtt{pre}(e)\}$,
- $(w,a) \trianglelefteq'_i (w',b)$ *iff one of the following clauses holds:*
  1. $a \trianglelefteq_i b$, $b \not\trianglelefteq a$ *and* $w \trianglelefteq w' \vee w' \trianglelefteq w$
  2. $a \trianglelefteq_i b$, $b \trianglelefteq a$ *and* $w \trianglelefteq w'$,
- $V'((s,e)) = V(s)$.

As mentioned in Section 2.9.3 our basic temporal doxastic agent properties in this setting are:

- $\trianglelefteq$-**Perfect Recall** : If $ha \trianglelefteq h'b$ we have $h \trianglelefteq h' \vee h' \trianglelefteq h$.
- $\trianglelefteq$-**Preference Propagation** : If $h \trianglelefteq h'$ and $ja \trianglelefteq j'b$ then also $ha \trianglelefteq h'b$.
- $\trianglelefteq$-**Preference Revelation** : If $ha \trianglelefteq h'b \wedge jb \trianglelefteq j'a$, also $h \trianglelefteq h'$.
- $\trianglelefteq$-**Accommodation** : If ($ja \trianglelefteq j'b$, $h' \trianglelefteq h$ and $ha \not\trianglelefteq h'b$), for all $ga, g'b \in H$ $(g \trianglelefteq g' \leftrightarrow ga \trianglelefteq g'b)$, and for all $g'a, gb \in H$ $(g \trianglelefteq g' \leftrightarrow gb \trianglelefteq g'a)$.

For simplicity we fix the precondition language here to be that of safe belief on local plausibility ordering. But for the reasons mentioned in Chapter 2 the result generalizes for more languages by adjusting the notion of bisimulation invariance.

Theorem C.1. *Let $\mathcal{H}$ be any local doxastic-temporal model with a local plausibility order. Then the following two assertions are equivalent:*

1. *There exists a local plausibility model $\mathcal{M}$ and a sequence of local plausibility event models $\vec{\epsilon}$ such that $\mathcal{H}$ is isomorphic to the forest generated by the Priority Update of $\mathcal{M}$ by the sequence $\vec{\epsilon}$.*

2. *$\mathcal{H}$ satisfies Propositional Stability, Synchronicity, $\trianglelefteq$-bisimulation invariance, $\trianglelefteq$-Perfect Recall, $\trianglelefteq$-Preference Propagation, $\trianglelefteq$-Preference Revelation and $\trianglelefteq$-Accommodation*

Proof. **Necessity** $(1 \implies 2)$**.** We show that the given conditions are satisfied by any local $DoTL$ model generated through successive priority updates along some given protocol sequence. *Propositional Stability* and *Synchronicity* are straightforward from the definition of generated forests.

$\trianglelefteq$-Preference Propagation. Assume that $ja \trianglelefteq j'b$; it follows by definition of local priority update that $a \trianglelefteq b$. Now assume $h \trianglelefteq h'$, it follows that in all cases ($b \trianglelefteq a$ or $b \not\trianglelefteq a$) we have by local priority update $ha \trianglelefteq h'b$.

$\trianglelefteq$-Preference Revelation. Assume that $ha \trianglelefteq h'b$, it follows that $a \trianglelefteq b$. Assuming $jb \trianglelefteq j'a$, we get $b \trianglelefteq a$. It follows from the definition of local priority update that $h \trianglelefteq h'$.

$\trianglelefteq$-Perfect Recall. Perfect recall is (almost) immediate from the definition of local priority update.

$\trianglelefteq$-Accommodation. Assume that $h' \trianglelefteq h$, $za \trianglelefteq z'b$ and $ha \not\trianglelefteq h'b$. Now assume for a contradiction that $a \triangleleft b$. It follows by local priority update that $ha \trianglelefteq h'b$, a contradiction. Now assume that $a \not\trianglelefteq b$, it follows by the definition of local priority update, that $za \not\trianglelefteq z'b$. But then $a \cong b$. By definition it is easy to show that the local plausibility relation is invariant under updates a and b, i.e. a and b are $\trianglelefteq$-accomodating.

Sufficiency $(2 \implies 1)$**.** Given a local $DoTL$ model $\mathcal{H}$ satisfying the stated conditions, we show how to construct a matching doxastic plausibility model and a sequence of event models.

Construction. Here is the initial plausibility model $\mathcal{M}_0 = \langle W, (\trianglelefteq_i^0)_{i \in N}, \hat{V} \rangle$:

- $W := \{h \in H \mid len(h) = 1\}$
- Define $h \trianglelefteq_i^0 h'$ whenever $h \trianglelefteq_i h'$.

- For every $p \in \text{PROP}$, define $\hat{V}(p) = V(p) \cap W$

We construct the j-th event model $\epsilon_j = \langle E_j, (\trianglelefteq_i^j)_{i \in N}, \mathtt{pre}_j \rangle$ as follows:

- $E_j := \{e \in \Sigma \mid \text{there is a history of the form } he \in H \text{ such that } len(h) = j\}$
- For each $i \in N$, define $a \trianglelefteq_i^j b$ iff
  - either there are $ha, h'b \in H$ such that $len(h) = len(h) = j$ and $ha \trianglelefteq_i h'b$.
  - or if for all g, g', $ga, g'b \in H$ we have $ga \trianglelefteq g'b$ iff $g \trianglelefteq g'$ and similarly when switching a and b — in other words if a and b always preserve and anti-preserve the previous order — in which case we put $a \cong b$ (i.e. $a \trianglelefteq b$ and $b \trianglelefteq a$). Whenever such a situation occurs, we say that a and b are *accommodating*. When this step is not involved in the construction of the plausibility order in an event model, we say that it is constructed in a normal way.

Equivalence: $h \trianglelefteq h'$ iff $h \trianglelefteq^F_{DDL} h'$. We now prove by induction on the length of the histories that $h \trianglelefteq h'$ iff $h \trianglelefteq^F_{DDL} h'$.

Assume that $len(h) = len(h') = 1$; then it is straightforward to see that plausibility is preserved and anti-preserved in the epistemic model by the construction, and therefore preserved in the generated ETL forest (again by construction).

Now for the induction step.

From $DoTL$ to $Forest(DEL)$

Case 1. $h \trianglelefteq h', ha \trianglelefteq h'b$ Assume $ha \trianglelefteq h'b$. By construction $a \trianglelefteq b$. Since $h \trianglelefteq h'$, by IH $h \trianglelefteq^F_{DDL} h'$ and thus at least one clause of local priority update applies and we have $ha \trianglelefteq^F_{DDL} h'b$.

Case 2. $h \not\trianglelefteq h', ha \trianglelefteq h'b$ Assume $ha \trianglelefteq h'b$ (1). It follows by construction that $a \trianglelefteq b$ (2). It also follows by $\trianglelefteq$-Perfect Recall that $h \trianglelefteq h' \vee h' \trianglelefteq h$, and by IH that $h \trianglelefteq^F_{DDL} h' \vee h' \trianglelefteq^F_{DDL} h$ (3). Moreover since by hypothesis $h \not\trianglelefteq h'$ (4) we have $h' \trianglelefteq h$ (5).

By (1) and (4) we know that a and b are not $\trianglelefteq$-accomodating (6). But by contraposition of $\trianglelefteq$-Preference Revelation, (1) and (4), we have that for all $j'a, jb \in H$ $jb \not\trianglelefteq j'a$ (7). But by (6) we know that only the normal part of the construction has been used, so that $b \not\trianglelefteq a$ (8).

Finally it follows from (1) that $a \trianglelefteq b$ (9). But by (3), (9), (8) and the first clause of local priority update it follows that $ha \trianglelefteq^F_{DDL} h'b$.

Converse direction: From $Forest(DEL)$ to $DoTL$

Case 1: $h \trianglelefteq^F_{DDL} h', ha \trianglelefteq^F_{DDL} h'b$**.** Assume that $h \trianglelefteq^F_{DDL} h'$; it follows by IH that $h \trianglelefteq h'$ (1). Now assume for a contradiction that $ha \not\trianglelefteq h'b$ (2). It follows that a and b are not accomodating and thus that the local plausibility relation is defined in a normal way. Now assume that $ha \trianglelefteq^F_{DDL} h'b$. It follows by the definition of local priority update that $a \trianglelefteq b$ and thus (by normality of the construction) that for some $ja, j'b \in H$, $j_a \trianglelefteq j'b$ but then by $\trianglelefteq$-Preference Propagation and (1) we have $ha \trianglelefteq h'b$, contradicting (2).

Case 2: $h \not\trianglelefteq^F_{DDL} h', ha \trianglelefteq^F_{DDL} h'b$**.** This is the only place where we make use of the abnormal part of the construction of the plausibility relation in the event models (and of axiom G^3).

Assume that $h \not\trianglelefteq^F_{DDL} h', ha \trianglelefteq^F_{DDL} h'b$. It follows by the definition of local priority update that $a \trianglelefteq b$ (1), $b \not\trianglelefteq a$ (2) and $h' \trianglelefteq^F_{DDL} h$ (3). From (2) it follows that a and b are not accomodating (4) (for otherwise the construction would give us $a \cong b$). But then by (1) and normality of the construction, it follows that there are some $ja, j'b \in H$ with $ja \trianglelefteq j'b$ (5).

Moreover by (3) and IH it follows that $h' \trianglelefteq h$ (6).

Now assume for a contradiction that $ha \not\trianglelefteq h'b$ (7). It follows from (5), (6), (7) and $\trianglelefteq$-Accommodation that a and b are accomodating, contradicting (4). Thus by reduction from (7) we have $ha \trianglelefteq h'b$. This concludes the proof for this case and the induction. QED

Appendix D

Additional proofs for Chapter 4

Proof of Fact 4.11 To show this result we introduce a few more definitions.

Definition D.1. *Two epistemic plausibility models $\mathcal{M}$ and $\mathcal{M}'$ are doxastically* bisimilar *whenever there is a relation $\underline{\leftrightarrow} \subseteq W \times W'$ such that for all $w \in W$ and $v \in W'$, if $w \underline{\leftrightarrow} v$ then:*

- *For all $p \in$ PROP, $w \in V(p)$ iff $v \in V'(p)$.*
- *Back and forth for $\sim_i$.*
  - *If $w \sim_i w'$ then there is a $v' \in W'$ such that $v \sim'_i v'$ and $w' \underline{\leftrightarrow} v'$.*
  - *If $v \sim_i v'$ then there is a $w \in W$ such that $w \sim'_i w'$ and $v' \underline{\leftrightarrow} w'$.*
- *For all formulas φ, back and forth for $\rightarrow_i^{||\varphi||}$. We write $\rightarrow_i^{\varphi}$ in what follows.*
  - *If $w \rightarrow_i^{\varphi} w'$ then there is a $v' \in W'$ such that $v \rightarrow_i'^{\varphi} v'$ and $w' \underline{\leftrightarrow} v'$.*
  - *If $v \rightarrow_i'^{\varphi} v'$ then there is a $w \in W$ such that $w \rightarrow_i^{\varphi} w'$ and $v' \underline{\leftrightarrow} w'$.*

Two pointed models $\mathcal{M}, w$ and $\mathcal{M}', v$ bisimilar, noted $\mathcal{M}, w \underline{\leftrightarrow} \mathcal{M}', v$, if $w \underline{\leftrightarrow} v$.

Fact D.2. *For all models $\mathcal{M}$ and $\mathcal{M}'$, $w \in W$, $v \in W'$ and $\varphi \in \mathcal{L}_{EDL}$, if $\mathcal{M}, w \underline{\leftrightarrow} \mathcal{M}', v$ then $\mathcal{M}, w \Vdash \varphi$ iff $\mathcal{M}', v \Vdash \varphi$.*

Proof of Fact 4.11 In the following proof we often write $[w]_1$ for $\mathcal{K}_1[w]$.

Proof. Let the model $\mathcal{M}$ be as in the proof of Theorem 4.9 and $\mathcal{M}'$ be defined as follows, with $I = \{1, 2\}$ in both cases. $\mathcal{M}' = \langle W, I, (\leq')_{i \in I}, (\sim')_{i \in I}, V' \rangle$ such that:

- $W = \{w_o, w_e\}$.
- $w_e <'_1 w_o$ and $w_o <'_2 w_e$.
- For both $i \in I$ and $w \in W$: $[w]_i = W$.

- $V'(p) = \{w_o\}$ and $V(q) = \emptyset$ for all $q \neq p$ in PROP.

Observation D.3. *In $\mathcal{M}$ there is a common prior and $\mathcal{M}, x \Vdash CB_I(B_1(\neg p) \wedge B_2(p))$ for all $x \in \mathbb{Z}$. In $\mathcal{M}'$ the latter is common belief as well, at both $w \in W$, but 1 and 2 have different priors.*

Define the relation $\underline{\leftrightarrow} \subseteq \mathcal{M} \times \mathcal{M}'$ as follows: $x \underline{\leftrightarrow} w_o$ for all odd integers x, and $x \underline{\leftrightarrow} w_e$ for all even integers.

Claim D.4. *The relation $\underline{\leftrightarrow}$ is a bisimulation.*

Proof. The propositional clause is trivial. It is easy to see that the clause for the relations $\sim_i$ and $\sim'_i$ is also satisfied. It remains to be shown that the clause for the families of relations $\rightarrow_i^{\varphi}$ and $\rightarrow_i'^{\varphi}$are also satisfied. We show this by induction on φ. In fact we show something stronger, namely that for all φ:

1. If $x \rightarrow_i^{\varphi} y$ and $x \underline{\leftrightarrow} w$ then there is a w' such that $w \rightarrow_i'^{\varphi} w'$ and $y \underline{\leftrightarrow} w'$.
2. If $w \rightarrow_i'^{\varphi} w'$ then for all $x \underline{\leftrightarrow} w$ there is a y such that $x \rightarrow_i^{\varphi} y$ and $w' \underline{\leftrightarrow} y$.
3. If x is odd then $x \in ||\varphi||^{\mathcal{M}}$ iff $w_o \in ||\varphi||^{\mathcal{M}'}$, and if x is even then $x \in ||\varphi||^{\mathcal{M}}$ iff $w_e \in ||\varphi||^{\mathcal{M}'}$.

Base Case: $\varphi \in$ PROP. We only have to consider p.

1. Assume that x is odd and that $x \rightarrow_i^p y$. Observe that by construction it can only be that $\min_{\leq_i}([x]_i \cap ||p||) = \{x\}$, for both $i = 1, 2$. This means that $y = x$, and so we are done, since $w_o \rightarrow_i'^p w_o$ and $x \underline{\leftrightarrow} w_o$. Suppose that x is even. Then $x \rightarrow_1^p y$ iff $y = x - 1$, again by construction. But since $y \underline{\leftrightarrow} w_o$ and $w_e \rightarrow_1'^p w_o$, we are done. The case for $x \rightarrow_2^p y$ is similar, with taking here $y = x + 1$.

2. Consider first w_o, and suppose that $w_o \rightarrow_i'^p w'$. Observe again that this can only happen if $w' = w_o$. Now take any x such that $w_0 \underline{\leftrightarrow} x$. By definition any such x is odd, and thus $x \in V(p)$. But we know, furthermore, that $\min_{\leq_i}([x]_i \cap ||p||) = \{x\}$ for both $i = 1, 2$, and so we are done. The case for w_e is entirely similar.

3. Follows directly from the definition of V and V'.

Inductive Step. Our inductive hypothesis is that claims (1), (2) and (3) hold for all φ' of lower complexity than φ. We only show the cases for (1): the arguments for (2) are entirely symmetrical, and the ones for (3) are simple applications of the inductive hypothesis.

- $\varphi := \neg\psi$.

1. We only show the case where x is odd. The other one is similar, with 1 and 2 reversed. Suppose that $x \rightarrow_i^{\neg\psi} y$. This means that $\mathcal{M}, y \nVdash \psi$.

Consider first the case where $x = y$. Then $\mathcal{M}, x \not\Vdash \psi$, and thus by our inductive hypothesis $\mathcal{M}', w_o \not\Vdash \psi$. This is enough to conclude that $w_o \rightarrow_2'^{\neg\psi} w_o$, simply because $w_e >_2' w_o$. So consider 1, for which we have that $[x]_1 = \{x, x+1\}$. Since $x >_1 x+1$, it must be that $\mathcal{M}, x+1 \Vdash \psi$. This means, by the inductive hypothesis again, that $\mathcal{M}', w_e \Vdash \psi$. But since $w_o >_1' w_e$, we have that $\min_{\leq_1'}([w_o]_1 \cap ||\neg\psi||) = \{w_o\}$, and so that $w_o \rightarrow_1'^{\neg\psi} w_o$. The reasoning for $x \neq y$ is similar, again with 1 and 2 reversed.

• $\varphi := \psi \wedge \xi$.

1. We show again only the case where x is odd. Suppose that $x \rightarrow_i^{\psi\wedge\xi} y$. Since $x \underline{\leftrightarrow} w_o$, we have to show that there is a $w' \in W$ such that $w_o \rightarrow_i'^{\psi\wedge\xi} w'$ and $y \underline{\leftrightarrow} w'$. Observe that either $y = x$ or $y = x+1$ if $i = 1$ and $y = x$ or $y = x-1$ if $i = 2$, which means that in both cases $y \in \min_{\leq_i}([x]_i \cap ||\psi|| \cap ||\xi||)$ iff

- either (*) $y \in \min_{\leq_i}([x]_i \cap ||\psi||)$, in which case, by the first clause of the inductive hypothesis, there is a w' such that $w_o \rightarrow_i'^{\psi} w'$ and $y \underline{\leftrightarrow} w'$;
- or (**) $y \in \min_{\leq_i}([x]_i \cap ||\xi||)$ in which case, again by the first clause of the inductive hypothesis, there is a w'' such that $w_o \rightarrow_i'^{\psi} w''$ and $y \underline{\leftrightarrow} w'$.

One can check that for each agent there is a unique $w' \in [w_o]_i$ such that $y \underline{\leftrightarrow} w'$, whatever y is, and so if both (*) and (**) hold then it must be that $w' = w''$, which means that we are done. We show the case where only (*) holds. For agent 1, this can only happen when $x = y$. By (*) and the inductive hypothesis this means that $w_o \rightarrow_i'^{\psi} w_o$, because there is no other $w' \in [w]_i$ such that $x \underline{\leftrightarrow} w'$. By assumption, we know furthermore that $\mathcal{M}, x \Vdash \xi$, and so by the inductive hypothesis that $\mathcal{M}', w_o \Vdash \xi$. It remains to be shown to $\mathcal{M}', w_e \not\Vdash \psi$. Since (**) does not hold, it has to be that $\mathcal{M}, x+1 \Vdash \xi$: we know that $\mathcal{M}, x \Vdash \xi$ and $x >_1 x+1$. This means that $\mathcal{M}, x+1 \not\Vdash \psi$, for otherwise we would have $x+1 \in \min_{\leq_1}([w]_i \cap ||\psi|| \cap ||\xi||)$, against the minimality of x. By the inductive hypothesis, then we know that $\mathcal{M}', w_e \not\Vdash \psi$, as required. The case for agent 2 follows the same line, except that (**) can only fail if $x \neq y$.

• $\varphi := K_i\psi$.

1. Suppose that x is odd and $x \rightarrow_i^{K_j\psi} y$. We only show the case for $i = 1$. Assume that $j = 1$ as well. Then $\mathcal{M}, y \Vdash K_1\psi$. By positive introspection of K_i, this means that $\mathcal{M}, y' \Vdash K_1\psi$ for all $y' \in [x]_1 = \{x, x+1\}$, and since $K_i\varphi \rightarrow \varphi$ is also valid for K_i, we get that $\mathcal{M}, y' \Vdash \psi$ for all such y'. By the inductive hypothesis this means that $\mathcal{M}', w_o \Vdash \psi$ and $\mathcal{M}', w_e \Vdash \psi$, and so that $\mathcal{M}', w_o \Vdash K_1\psi$ and $\mathcal{M}', w_e \Vdash K_1\psi$. Since y is either x or $x+1$, we get that for any $w' \underline{\leftrightarrow} y$, $w_o \rightarrow_1'^{K_1\psi} w'$. Suppose now that $j = 2$ and $y = x$. This means that $\mathcal{M}, x \Vdash K_2\psi$. Again by positive introspection

and the truth axiom, we get that $\mathcal{M}, x \Vdash \psi$ and $\mathcal{M}, x-1 \Vdash \psi$. Using our inductive hypothesis twice, we conclude that $\mathcal{M}', w_o \Vdash \psi$ and $\mathcal{M}', w_e \Vdash \psi$. But this covers all $w' \in [w_o]_2 = [w_o]_1$, and in particular w_o, so we have that $w_o \rightarrow_1'^{K_2\psi} w_o$. The same reasoning applies *mutatis mutandis* when $y = x+1$, and if x is even.

- $\varphi := B_j^\xi \psi$.

1. Suppose that x is odd and $x \rightarrow_i^{B_j^\xi \psi} y$. Assume that $i \neq j$, and suppose that $i = 1$, the argument for $i = 2$ being entirely symmetric. We first show that it cannot be the case that $y = x$, for it would imply that $\mathcal{M}', w_o \Vdash B_2^\xi \psi$ while $\mathcal{M}', w_e \Vdash \neg B_2^\xi \psi$, which is impossible since $[w_o]_2 = [w_e]_2$. If $x = y$ then by the minimality of x within $[x]_1 \cap ||B_2^\xi \psi||$ it must be that both:

- (*) $\mathcal{M}, x \Vdash B_2^\xi \psi$ and
- (**) $\mathcal{M}, x+1 \Vdash \neg B_2^\xi \psi$.

If (*) then for all $y' \in \min_{\leq_2}([x]_2 \cap ||\xi||)$ we have that $\mathcal{M}, y' \Vdash \psi$. If $[x]_2 \cap ||\xi||$ is empty, then $\mathcal{M}, x \nVdash \xi$ and $\mathcal{M}, x \nVdash \xi$, which means by our inductive hypothesis that $\mathcal{M}', w_o \nVdash \xi$ and $\mathcal{M}', w_e \nVdash \xi$, and so that $\mathcal{M}', w_o \Vdash B_2^\xi \psi$ trivially. If $x \in \min_{\leq_2}([x]_2 \cap ||\xi||) \subseteq ||\psi||$, then we know by the inductive hypothesis that $\mathcal{M}, w_o \Vdash \xi \wedge \psi$. But since w_o is $\leq_2'$-minimal in $[w_o]_2$, we can conclude that $\mathcal{M}', w_o \Vdash B_2^\xi \psi$ as well. Finally, if $x \notin \min_{\leq_2}([x]_2 \cap ||\xi||) \subseteq ||\psi||$, it must be that $x \rightarrow_2^\xi x-1$. By the inductive hypothesis we know that $w_o \rightarrow_2^\xi w_e$, which can only happen if $\mathcal{M}', w_o \Vdash \neg\xi$, from which we can conclude that $\mathcal{M}', w_o \Vdash B_2^\xi \psi$. So from (*) we get that $\mathcal{M}', w_o \Vdash B_2^\xi \psi$. Now, by (**) we know that there is a $y' \in \min_{\leq_2}([x+1] \cap ||\xi||)$ such that $x \notin ||\psi||$. This y' is either $x+1 \leftrightarrow w_e$ or $x+2 \leftrightarrow w_o$. In the first case we get from our inductive hypothesis that $w_e \rightarrow_2'^{\xi} w_e$, which can only happen if $\min_{\leq_2}([x+1] \cap ||\xi||) = \{w_e\}$, and thus if $\mathcal{M}', w_e \Vdash \neg B_2^\xi \psi$. In the second case we get by the inductive hypothesis that $\mathcal{M}', w_o \Vdash \xi \wedge \neg\psi$, from which we also know that $\mathcal{M}', w_e \Vdash \neg B_2^\xi \psi$, since w_o is $\leq_2'$-minimal in $[w_o]_2$ and $B_i^\varphi \varphi' \rightarrow K_i B_i^\varphi \varphi'$ is valid in epistemic plausibility models. From (**) we thus know that $\mathcal{M}', w_e \Vdash \neg B_2^\xi \psi$, which means in conjunction with (*) that it cannot be that $x = y$.

Assume thus that $x \neq y$. This means that $x \rightarrow_1^{B_2^\xi \psi} x+1$. We are done if we can show that $w_o \rightarrow_i'^{B_j^\xi \psi} w_e$, for which it is enough to show that $\mathcal{M}', w_e \Vdash B_2^\xi \psi$. That $x \rightarrow_1^{B_2^\xi \psi} x+1$ means that $\mathcal{M}, x_1 \Vdash B_2^\xi \psi$. From there we reach the intended conclusion by following the same steps as above for (*).

Suppose then that $i = j = 1$. Then $x \rightarrow_1^{B_1^\xi \psi} y$. This means that $\mathcal{M}, y \Vdash B_1^\xi \psi$ and $\mathcal{M}, x \Vdash B_1^\xi \psi$ because $B_i^\varphi \varphi' \rightarrow K_i B_i^\varphi \varphi'$ is valid in epistemic plausibility

models. This means, first, $x \rightarrow_1^{\top} y$ and thus by the inductive hypothesis that there is a $w' \underline{\leftrightarrow} y$ such that $w_o \rightarrow_1'^{\top} w'$, i.e. that w' is $\leq_1'$-minimal in $[w_o]_i$. If we can show that $\mathcal{M}', w_o \Vdash B_1^{\xi}\psi$ then we are thus done. If $\mathcal{M}, x \Vdash B_1^{\xi}\psi$ because $[x]_i \cap ||\xi|| = \emptyset$ then we are done. Otherwise, if $\mathcal{M}, x+1 \Vdash \xi$ then by the inductive hypothesis we know that $\mathcal{M}', w_e \Vdash \xi \wedge \psi$ and so we are done because w_e is $\leq_1'$-minimal in $[w_o]_i = [w_e]_i$. If finally $\mathcal{M}, x+1 \nVdash \xi$ but yet $[x]_i \cap ||\xi|| \neq \emptyset$ then it must be that $\mathcal{M}, x \Vdash \xi \wedge \psi$. But then by the inductive hypothesis we know that $\mathcal{M}', w_o \Vdash \xi \wedge \psi$ and $\mathcal{M}', w_e \nVdash \xi$, which is enough to show that $\mathcal{M}', w_o \Vdash B_1^{\xi}\psi$. The argument for $i = j = 2$ is symmetric.

- $\varphi := C_G\psi$.

1. The case when G is a singleton boils down to knowledge. So we consider the case were $G = \{1, 2\}$. Assume $x \underline{\leftrightarrow} w$ (H) and $x \rightarrow_i^{C_{\{1,2\}}\psi} y$ (0). By definition of $\rightarrow_i^{C_{\{1,2\}}\psi}$ it follows that $\mathcal{M}, y \Vdash C_{\{1,2\}}\psi$. By definition of $\mathcal{M}$ it follows that for all $z \in |\mathcal{M}| = \mathbb{Z}$ we have $\mathcal{M}, z \Vdash \psi$. By IH it follows that for all $v \in |\mathcal{M}'|$ we have $\mathcal{M}', v \Vdash \psi$. Moreover in both models $C_{\{1,2\}}\psi$ is satisfied everywhere (1). Now assume that $i = 1$, and that x is even. It follows that $\mathcal{K}_1[x] = \{x-1, x\}$. Moreover x is the minimum of $\mathcal{K}_1[x]$ (2). So by (0), (1) and (2) we have $x = y$. Now we have $x \underline{\leftrightarrow} w_e$ and in the second model $w_e \rightarrow_i^{C_{\{1,2\}}\psi} w_e$. Now assume that x is odd. It follows that $\mathcal{K}_1[x] = \{x, x+1\}$. Moreover $x+1$ is the minimum of $\mathcal{K}_1[x]$ (3). So by (0), (1) and (3) we have $x+1 = y$. Now we have $y = x+1 \underline{\leftrightarrow} w_e$ and in the second model $w_o \rightarrow_i^{C_{\{1,2\}}\psi} w_e$. Now assume $i = 2$ and that x is odd. It follows that $\mathcal{K}_2[x] = \{x-1, x\}$. Moreover x is the minimum of $\mathcal{K}_2[x]$ (2). So by (0), (1) and (2) we have $x = y$. Now we have $x \underline{\leftrightarrow} w_o$ and in the second model $w_o \rightarrow_i^{C_{\{1,2\}}\psi} w_o$. Suppose finally that x is even. It follows that $\mathcal{K}_2[x] = \{x, x+1\}$. Moreover $x+1$ is the minimum of $\mathcal{K}_2[x]$ (3). So by (0), (1) and (3) we have $x+1 = y$. Now we have $y = x+1 \underline{\leftrightarrow} w_o$ and in the second model $w_e \rightarrow_i^{C_{\{1,2\}}\psi} w_o$.

- $\varphi := CB_G\psi$.

We have that $x \rightarrow_i^{CB_G\varphi} y$ iff $x \rightarrow_i^{C_G\varphi} y$ in $\mathcal{M}$, and similarly in $\mathcal{M}'$.

QED

This concludes the proof of the Claim and the whole argument. QED

Appendix E

Additional proofs for Chapter 5

In this appendix we give the proofs we omitted in Chapter 5 concerning multi-agent learning. The main step to prove Theorem 5.29 will be to prove a No Learning Theorem that can be stated as follows:

Theorem E.1. *Whenever agents have the same background information and the same observational powers, then there is no knowledge gain by forcing announcement of conjectures between each step.*

Formally, let $\mathcal{M}$ be a doxastic epistemic model that satisfies same initial information and P a local protocol for $\mathcal{M}$ in which all agents have the same observational capacities. Let $\mathcal{F}or(\mathcal{M},P)[(\sim_i)_{i\in N}]$ be the doxastic epistemic forest generated by $\mathcal{M}$ and P and $\mathcal{F}or(\mathcal{M},!B(p))[(\sim_i^{!B})_{i\in N}]$ be the doxastic epistemic forest generated by $\mathcal{M}$ and $!B(p)$. We have:

$$h \sim_i h' \text{ iff } !B(h) \sim_i^{!B} !B(h') \text{ for all } i \in N.$$

Now for the proof of this theorem.

Proof of Theorem E.1. In the proof of this theorem we will need the following fact and the two following lemmas.

Fact E.2 (DETL models satisfy belief introspection)**.**
If $K_i[h] = K_i[h']$ then $B_i[h] = B_i[h']$.

Proof. $B_i[h] = \min_{\leq_i} K_i[h] = \min_{\leq_i} K_i[h'] = B_i[h']$. QED

Lemma E.3 (Same Info Lemma)**.** *Let $\mathcal{H}=\langle W, \Sigma, H, (\leq_i)_{i\in N}, (\sim_i)_{i\in N}, V\rangle$ be a doxastic epistemic model satisfying SOC(i,j), PR(i,j), SYN(i,j) and SII(i,j). It follows that for all $h', h \in H$, we have $h \sim_i h'$ iff $h \sim_j h'$.*

Proof. The proof is by induction on the length of h, h'. The proof by induction is justified by *SYN*(i,j).
Base case is immediate by *SII*(i,j).

Induction step. Assume that $ve_1 \dots e_{n+1} \sim_i we'_1 \dots e'_{n+1}$ (a). By $PR(i)$ we have $ve_1 \dots e_n \sim_i we'_1 \dots e'_n$ (b). But then by IH we have $ve_1 \dots e_n \sim_j we'_1 \dots e'_n$ (c). From (b), (c), (a) and $SOC(i,j)$ it follows that $ve_1 \dots e_{n+1} \sim_j we'_1 \dots e'_{n+1}$. The other direction is of course identical. QED

Lemma E.4 (Inter-model No Miracles Lemma)**.** *Let $\mathcal{M}$ be a doxastic epistemic model. Let P be a local protocol for $\mathcal{M}$. Let $\mathcal{F}or(\mathcal{M}, p) = \langle W, \Sigma, H, (\leq_i)_{i\in N}, (\sim_i)_{i\in N}, V\rangle$ be the doxastic epistemic forest generated by $\mathcal{M}$ and P and $\mathcal{F}or(\mathcal{M}, !B(p)) = \langle W, \Sigma \cup \{!B\}, H^{!B}, (\leq_i)_{i\in N}, (\sim_i^{!B})_{i\in N}, V\rangle$ be the doxastic epistemic forest generated by $\mathcal{M}$ and $!B(P)$. If $we_1 \dots e_{n+1} \sim_i ve'_1 \dots e'_{n+1}$ and $w!B(e_1 \dots e_n) \sim_i^{!B} v!B(e'_1 \dots e'_n)$, then $w!B(e_1 \dots e_n)e_{n+1} \sim_i^{!B} v!B(e'_1 \dots e'_n)e_{n+1}$.*

Proof. By hypothesis $we_1 \dots e_{n+1} \sim_i ve'_1 \dots e'_{n+1}$. But then by the definition of product update we have $e_{n+1} \sim_i e'_{n+1}$. By definition of $!B(P)$ it follows that $e_{n+1} \sim_i^{!B} e'_{n+1}$ (1). Now by hypothesis we have $w!B(e_1 \dots e_n) \sim_i^{!B} v!B(e'_1 \dots e'_n)$ (2). But then by (1), (2) and product update it follows that $w!B(e_1 \dots e_n)e_{n+1} \sim_i^{!B} v!B(e'_1 \dots e'_n)e_{n+1}$. QED

We can now start with the proof of Theorem E.1.

Proof. The proof is by induction on the length of the history, which is allowed by Synchronicity.[1] We start with the easy direction: from right to left.
Base case Assume that $!B(w) \sim_i^{!B} !B(v)$. It follows by perfect recall that $w \sim_i^{!B} v$. But by construction the initial models are identical, thus $w \sim_i v$.
Induction step. Assume that $w!B(e_1 \dots e_n e_{n+1}) \sim_i^{!B} v!B(e'_1 \dots e'_n e'_{n+1})$. It follows by Perfect Recall that $w!B(e_1 \dots e_n)e_{n+1} \sim_i^{!B} v!B(e'_1 \dots e'_n)e'_{n+1}$ (1). By Perfect Recall, it follows from (1) that $w!B(e_1 \dots e_n) \sim_i^{!B} v!B(e'_1 \dots e'_n)$ (2). By IH and (2) we have $we_1 \dots e_n \sim_i ve'_1 \dots e'_n$ (3). But then by Lemma E.4, (1) and (3) it follows that $we_1 \dots e_{n+1} \sim_i ve'_1 \dots e'_{n+1}$.

Now for the more interesting direction: From left to right. We re-start counting of propositions.
Base case. Assume that $v \sim_i w$ (1). We prove that $v!B \sim_i^{!B} w!B$. Take an arbitrary agent j. From (1) we have $K_i[v] = K_i[w]$ (2). By $SII(i,j)$ and (2) it follows that:

$$K_j[w] = K_i[w] = K_i[v] = K_j[v] \quad (3)$$

But then by belief introspection for j (Fact E.2) and (3) we have $B_j[w] = B_j[v]$ (4). Since j was arbitrary it follows from (4), (2) "$!B$ is the *Belief Announcement* event" that $w!B \sim_i^{!B} v!B$.

Induction step. Assume that $ve_1 \dots e_{n+1} \sim_i we'_1 \dots e'_{n+1}$ (5). We prove that $w!B(e_1 \dots e_{n+1}) \sim_i^{!B} v!B(e'_1 \dots e'_{n+1})$. First of all it follows from (5) and perfect

[1] In the following proof the usage of properties such as Synchronicity is justified by Corollary 5.16 when talking about $\mathcal{F}or(\mathcal{M}, p)$ and by another easy corollary, whose proof we omit, when talking about $\mathcal{F}or(\mathcal{M}, !B(p))$. We drop further reference to these two results in the proof.

recall for i that $ve_1 \dots e_n \sim_i we'_1 \dots e'_n$ (6). But then by (6) and IH we have $w!B(e_1 \dots e_n) \sim_i^{!B} v!B(e'_1 \dots e'_n)$ (7).

Now take an arbitrary $j \in N$. By (6) and Lemma E.3 it follows that $ve_1 \dots e_n \sim_j we'_1 \dots e'_n$ (8). By IH and (8) we have $v!B(e_1 \dots e_n) \sim_j^{!B} w!B(e'_1 \dots e'_n)$ (9). By (5) and Lemma E.3 we have $ve_1 \dots e_{n+1} \sim_j we'_1 \dots e'_{n+1}$ (10).

Now from (10), (9) and uniform no miracles for j (for all events in Σ) it follows that $v!B(e_1 \dots e_n)e_{n+1} \sim_j^{!B} w!B(e'_1 \dots e_n)e'_{n+1}$ (11). Similarly from (5), (7) and uniform no miracles for i (for all events in Σ) we have $v!B(e_1 \dots e_1)e_{n+1} \sim_j^{!B} w!B(e'_1 \dots e_n)e'_{n+1}$ (12). By belief introspection for j and (11) it follows that $B_j[v!B(e_1 \dots e_1)e_{n+1}]{=}B_j[w!B(e'_1 \dots e_n)e'_{n+1}]$ (13). Since j was arbitrary it follows from (13), (12) and "$!B$ is the *Belief Announcement* event" that $v!B(e_1 \dots e_{n+1}) \sim_i^{!B} w!B(e'_1 \dots e_n e'_{n+1})$ QED

Proof of Theorem 5.29 We start by proving the following corollary which will make it very easy to prove Theorem 5.29.

Corollary E.5. *let $\mathcal{M}$ be a doxastic epistemic model that satisfies SII(N) and P a local protocol for $\mathcal{M}$ such that P satisfies SOC(N). Let $\mathcal{F}or(\mathcal{M}, p) = \langle W, \Sigma, H, (\leq_i)_{i \in N}, (\sim_i)_{i \in N}, V \rangle$ be the doxastic epistemic forest generated by $\mathcal{M}$ and p and $\mathcal{F}or(\mathcal{M}, !B(p)) = \langle W, \Sigma \cup \{!B\}, H^{!B}, (\leq_i)_{i \in N}, (\sim_i^{!B})_{i \in N}, V \rangle$ be the doxastic epistemic forest generated by $\mathcal{M}$ and $!B(p)$.*

We have $\exists h = wh' \in B_i[ve_1 \dots e_n]$ iff $\exists h_2 = wh_3 \in B_i^{!B}[!B(ve_1 \dots e_n)]$.

Proof. The proof is by induction on the length of $ve_1 \dots e_n$ which is allowed by the assumption of Synchronicity.[2]

Base case. We prove both directions simultaneously. By construction we have for all $v, w \in W$ $w \leq_i v$ iff $w!B \leq_i v!B$. Since by Theorem E.1 we have $w \sim_i v$ iff $w!B \sim_i v!B$. It follows that for all $v, w \in W$ we have $w \in B_i[v] = \min_{\leq_i} K_i[v]$ iff $w!B \in B_i[v!B] = \min_{\leq_i} K_i[v!B]$.

Induction step. From left to right. Assume that there is some history $wh \in B_i[ve_1 \dots ve_{n+1}]$ (1). It follows that $wh \in K_i[ve_1 \dots ve_{n+1}]$, i.e. $wh \sim_i ve_1 \dots ve_{n+1}$ (2). But then by Theorem E.1 we have $w!B(h) \sim_i v!B(e_1 \dots ve_{n+1})$ (3). Now assume for a contradiction that for every h' of the form $w!B(h_2)$, we have $h' \notin B_i[!B(ve_1 \dots e_{n+1})]$ (4). It follows that for every such h' we have some history $s!B(h_3) \in K_i[!B(ve_1 \dots e_{n+1})]$ (5) with $s \neq w$ (6) and $s!B(h_3) <_i h'$ (7). It is easy to check that $len(sh_3) = len(h')$ (8). But then by (7), (8) and Preference Stability we have $s <_i^{!B} w$ (9). By construction it follows that $s < w$ (10). But by (5) and Theorem E.1 we have $sh_3 \in K_i[ve_1 \dots e_{n+1}]$ (11). But then by Preference Stability we have $sh_3 < w(h_2)$ (12). Thus, by definition of B_i, we have $w(h_2) \notin B_i[ve_1 \dots ve_{n+1}]$ (13). But since h_2 was arbitrary we have in particular $wh \notin B_i[ve_1 \dots ve_{n+1}]$ (14), contradicting (1). Thus by reduction there is some history h' of the form $w!B(h_2)$ such that $h' \in B_i[!B(ve_1 \dots e_{n+1})]$ (15).

[2]Same remark as for the previous proof.

The other direction is similar. QED

The proof of Theorem 5.29 is now easy.

Proof. Assume that i stabilizes on $v \in W$ after the sequence $we_1 \dots e_n$. Then for every h which extends $we_1 \dots e_n$, all histories in $B_i[h]$ starts with v. But then for every h, by Corollary E.5, at $!B(h)$, i believes only in histories starting with v. The other direction is similar. QED

Bibliography

[1] Thomas Ågotnes, Paul E. Dunne, Wiebe van der Hoek, and Michael Wooldridge. Logics for coalitional games. In Johan van Benthem, Shier Ju, and Frank Veltman, editors, *A Meeting of the Minds – Proceedings of the Workshop on Logic, Rationality and Interaction, Beijing, 2007*, pages 3–20, London, 2007. College Publications.

[2] Thomas Ågotnes, Wiebe van der Hoek, and Michael Wooldridge. Quantified coalition logic. In M. M. Veloso, editor, *Proceedings of the Twentieth International Joint Conference on Artificial Intelligence (IJCAI 2007)*, pages 1181–1186, California, 2007. AAAI Press.

[3] Carlos E. Alchourrón, Pater Gärdenfors, and David Makinson. On the logic of theory change: partial meet contraction and revision functions. *Journal of Symbolic Logic*, 50(2):510–530, 1985.

[4] Rajeev Alur, Thomas A. Henzinger, and Orna Kupferman. Alternating-time temporal logic. In *Proceedings of the 38th IEEE Symposium on Foundations of Computer Science*, Florida, October 1997.

[5] Dana Angluin. Inductive inference of formal languages from positive data. *Information and Control*, 45(2):117–135, 1980.

[6] Krzysztof R. Apt. Epistemic analysis of strategic games with arbitrary strategy sets. In Dov Samet, editor, *Proceedings 11th Conference on Theoretical Aspects of Reasoning about Knowledge (TARK07)*, pages 25–32. PUL, 2007.

[7] 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.

[8] Carlos Areces, Patrick Blackburn, and Maarten Marx. A road-map on complexity for hybrid logics. In Flum and Rodríguez-Artalejo, editors, *CSL*, number 1683 in LNCS, pages 307–321, Madrid, Spain, 1999. Springer. Proceedings of the 8th Annual Conference of the EACSL.

[9] Carlos Areces, Patrick Blackburn, and Maarten Marx. Hybrid logics: characterization, interpolation and complexity. *The Journal of Symbolic Logic*, 66(3):977–1010, 2001.

[10] Robert J. Aumann. Backward induction and common knowledge of rationality. *Games and Economic Behavior*, 8:6–19, 1995.

[11] Robert J. Aumann. Interactive epistemology i: Knowledge. *International Journal of Game Theory*, 28(3):263–300, 1999.

[12] Robert J. Aumann. Interactive epistemology ii: Probability. *International Journal of Game Theory*, 28(3):301–314, 1999.

[13] Robert J. Aumann. Agreeing to disagree. *The Annals of Statistics*, 4(6): 1236–1239, 1976.

[14] Michael Bacharach. Some extensions of a claim of Aumann in an axiomatic model of knowledge. *Journal of Economic Theory*, 37(1):167–190, October 1985.

[15] Alexandru Baltag and Lawrence S. Moss. Logics for Epistemic Programs. *Synthese*, 139(2):165–224, 2004.

[16] Alexandru Baltag and Sonja Smets. Dynamic belief revision over multi-agent plausibility models. In G. Bonanno, W. van der Hoek, and M. Wooldridge, editors, *Logic and the Foundations of Game and Decision Theory: Proceedings of LOFT'06*, TLG, pages 11–24. AUP, 2006.

[17] 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 Foundation of Game and Decision Theory (LOFT7)*, volume 3 of *Texts in Logic and Games*, pages 13–60. Amsterdam University Press, 2008.

[18] Alexandru Baltag and Sonja Smets. Group belief dynamics under iterated revision: fixed points and cycles of joint upgrades. In Heifetz [100], pages 41–50.

[19] Alexandru Baltag and Sonja Smets. Conditional doxastic models: A qualitative approach to dynamic belief revision. In G. Mints and R. de Queiroz, editors, *Proceedings of WOLLIC 2006, Electronic Notes in Theoretical Computer Science*, volume 165, 2006.

[20] Alexandru Baltag, Lawrence S. Moss, and Slawomir Solecki. The logic of public announcements, common knowledge, and private suspicions. In *TARK '98: Proceedings of the 7th conference on Theoretical Aspects of Rationality and Knowledge*, pages 43–56, San Francisco, CA, USA, 1998. Morgan Kaufmann Publishers Inc.

[21] 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.

[22] Pierpaolo Battigalli. On rationalizability in extensive games. *Journal of Economic Theory*, 74(1):40–61, 1997.

[23] Pierpaolo Battigalli and Giacomo Bonanno. Synchronic information, knowledge and common knowledge in extensive games. *Research in Economics*, 53(1):77 – 99, 1999.

[24] Nuel Belnap, Michael Perloff, and Ming Xu. *Facing the future: Agents and choices in our indeterminist world*. Oxford University Press, Oxford, 2001.

[25] Elchanan Ben-Porath. Rationality, Nash equilibrium and backwards induction in perfect information games. *The Review of Economic Studies*, 64(1): 23–46, 1997.

[26] Johan van Benthem. Extensive games as process models. *Journal of Logic, Language and Information*, 11:289–313(25), 2002.

[27] Johan van Benthem. *Modal Logic and Classical Logic*. Bibliopolis, Napoli, 1983.

[28] Johan van Benthem. Games in Dynamic Epistemic Logic. *Bulletin of Economic Research*, 53(4):219–248, 2001.

[29] Johan van Benthem. Dynamic logic for belief change. *Journal of Applied Non-Classical Logics*, 17(2):129–155, 2007.

[30] Johan van Benthem. Man muss immer umkehren. In Cédric Dégremont, Laurent Keiff, and Helge Rückert, editors, *Dialogues, Logics and other strange things. Essays in honour of Shahid Rahman*, Tributes, pages 53–65. College Publications, London, 2008.

[31] Johan van Benthem. Rational dynamics and epistemic logic in games. *International Game Theory Review*, 9(1):377–409, 2007. Erratum reprint, Volume 9(2), 2007, 377–409.

[32] Johan van Benthem and Cédric Dégremont. Multi-agent belief dynamics: bridges between dynamic doxastic and doxastic temporal logics. In Giacomo Bonanno, Wiebe van der Hoek, and Benedikt Löwe, editors, *Post-proceedings of the 8th Conference on Logic and the Foundations of Game and Decision Theory (LOFT08)*, Texts in Logic and Games. Amsterdam University Press, to appear.

[33] Johan van Benthem and Daisuke Ikegami. Modal fixed-point logic and changing models. In A. Avron, N. Dershowitz, and A. Rabinovich, editors, *Pillars of Computer Science*, volume 4800 of *Lecture Notes in Computer Science*, pages 146–165. Springer, 2008.

[34] Johan van Benthem and Eric Pacuit. The Tree of Knowledge in Action: Towards a Common Perspective. In I. Hodkinson G. Governatori and Y. Venema, editors, *Advances in Modal Logic*, volume 6. College Publications, 2006.

[35] Johan van Benthem, Jan van Eijck, and Barteld P. Kooi. Logics of communication and change. *Information and Computation*, 204(11):1620–1662, 2006.

[36] Johan van Benthem, Jelle Gerbrandy, Tomohiro Hoshi, and Eric Pacuit. Merging frameworks for interaction: DEL and ETL. *Journal of Philosophical Logic*, 2009.

[37] Johan van Benthem. One is a lonely number. In Z. Chatzidakis, P. Koepke, and W. Pohlers, editors, *Logic Colloquium '02*, Wellesley MA, 2006. ASL & A.K. Peters.

[38] Ken Binmore. A note on backward induction. *Games and Economic Behavior*, 17(1):135–137, 1996.

[39] Patrick Blackburn, Maarten de Rijke, and Yde Venema. *Modal Logic*. Number 53 in Cambridge Tracts in Theoretical Computer Science. Cambridge University Press, UK, 2001.

[40] Oliver Board. Dynamic interactive epistemology. *Games and Economic Behavior*, 49:49–80, 2004.

[41] Giacomo Bonanno. Prediction in branching time logic. *Math. Log. Q.*, 47 (2):239–247, 2001.

[42] Giacomo Bonanno. Revising predictions. In *TARK '01: Proceedings of the 8th conference on Theoretical Aspects of Rationality and Knowledge*, pages 273–286, San Francisco, CA, USA, 2001. Morgan Kaufmann Publishers Inc.

[43] Giacomo Bonanno. A syntactic approach to rationality in games with ordinal payoffs. In G. Bonanno, W. van der Hoek, and M. Wooldridge, editors, *Logic and the Foundation of Game and Decision Theory (LOFT7)*, volume 3 of *TLG*, pages 59–86. AUP, 2008.

[44] Giacomo Bonanno. Memory and perfect recall in extensive games. *Games and Economic Behavior*, 47(2):237 – 256, 2004.

[45] Giacomo Bonanno. Players' information in extensive games. *Mathematical Social Sciences*, 24(1):35–48, 1992.

[46] Giacomo Bonanno. Branching time, perfect information games, and backward induction. *Games and Economic Behavior*, 36:57–73, 2001.

[47] Giacomo Bonanno. Axiomatic characterization of the AGM theory of belief revision in a temporal logic. *Artif. Intell.*, 171(2–3):144–160, 2007.

[48] Giacomo Bonanno and Klaus Nehring. How to make sense of the common prior assumption under incomplete information. *International Journal of Game Theory*, 28(03):409–434, August 1999.

[49] Giacomo Bonanno and Klaus Nehring. Agreeing to disagree: a survey. Some of the material in this paper was published in [48], 1997.

[50] Nicolas Bourbaki. Sur le théorème de Zorn. *Archiv der Mathematik*, 2, 1949.

[51] Adam Brandenburger and H. Jerome Keisler. An impossibility theorem on beliefs in games. *Studia Logica*, 84(2):211–240, November 2006.

[52] Jordi Brandts and Charles A Holt. An experimental test of equilibrium dominance in signaling games. *American Economic Review*, 82(5):1350–65, December 1992.

[53] Jan Broersen, Andreas Herzig, and Nicolas Troquard. Normal Coalition Logic and its conformant extension. In Dov Samet, editor, *TARK'07*, pages 91–101. PUL, 2007. URL `http://www.irit.fr/~Andreas.Herzig/P/Tark07ncl.html`.

[54] Colin F. Camerer. *Behavioral Game Theory: Experiments in Strategic Interaction*. Princeton: Princeton University Press, 2003.

[55] Balder ten Cate. *Model theory for extended modal languages*. PhD thesis, University of Amsterdam, 2005. ILLC Dissertation Series DS-2005-01.

[56] Balder ten Cate and Massimo Franceschet. On the complexity of hybrid logics with binders. In L. Ong, editor, *Proceedings of Computer Science Logic 2005*, volume 3634 of *Lecture Notes in Computer Science*, pages 339–354. Springer Verlag, 2005.

[57] Jonathan A. K. Cave. Learning to agree. *Economics Letters*, 12(2):147–152, 1983.

[58] Brian F. Chellas. *Modal Logic: an introduction*. Cambridge University Press, 1980.

[59] Anuj Dawar, Erich Grädel, and Stephan Kreutzer. Inflationary fixed points in modal logic. *ACM Trans. Comput. Log.*, 5(2):282–315, 2004.

[60] Dick de Jongh and Fenrong Liu. Optimality, belief and preference. In Sergei Artemov and Rohit Parikh, editors, *Proceedings of the Workshop on Rationality and Knowledge*. ESSLLI, Malaga, 2006.

[61] Cédric Dégremont and Nina Gierasimczuk. Can doxastic agents learn? On the temporal structure of learning. In He et al. [99], pages 90–104.

[62] Cédric Dégremont and Lena Kurzen. Modal logics for preferences and cooperation: Expressivity and complexity. In *Knowledge Representation for Agents and Multi-Agent Systems, Post-proceedings of KRAMAS 2008*, volume 5605 of *Lecture Notes in Artificial Intelligence*. Springer-Verlag, 2009.

[63] Cédric Dégremont and Lena Kurzen. Getting together: A unified perspective on modal logics for coalitional interaction. In He et al. [99], pages 317–318.

[64] Cédric Dégremont and Olivier Roy. Agreement theorems in dynamic-epistemic logic. In Heifetz [100], pages 91–98.

[65] Cédric Dégremont and Jonathan A. Zvesper. Logique dynamique pour le raisonnement stratégique dans les jeux extensifs. In Jérôme Lang, Yves Lespérance, David Sadek, and Nicolas Maudet, editors, *Journées Francophones sur les Modèles Formels de l'Interaction (MFI'07)*, pages 61–73. Lamsade, 2007.

[66] Hans van Ditmarsch. Prolegomena to dynamic logic for belief revision. *Synthese*, 147:229–275, 2005.

[67] Hans van Ditmarsch, Wiebe van der Hoek, and Barteld Kooi. *Dynamic Epistemic Logic*, volume 337 of *Synthese Library Series*. Springer, 2007.

[68] Francesco M. Donini, Maurizio Lenzerini, Daniele Nardi, and Werner Nutt. The complexity of concept languages. In *KR*, pages 151–162, 1991.

[69] Jan van Eijck and Martin Stokhof. The gamut of dynamic logics. In Dov Gabbay and John Woods, editors, *Logic and the Modalities in the Twentieth Century*, volume 6 of *The Handbook of History and Philosophy of Logic*. Elsevier, 2005.

[70] Kfir Eliaz and Efe A. Ok. Indifference or indecisiveness? Choice-theoretic foundations of incomplete preferences. *Games and Economic Behavior*, 56 (1):61–86, 2006.

[71] Ronald Fagin, Joseph Y. Halpern, and Moshe Y. Vardi. What can machines know? On the properties of knowledge in distributed systems. *J. ACM*, 39 (2):328–376, 1992.

[72] Ronald Fagin, Joseph Y. Halpern, Yoram Moses, and Moshe Y. Vardi. *Reasoning About Knowledge*. MIT Press, Cambridge, 1995. ISBN 0262061627.

[73] Solomon Feferman. Persistent and invariant formulas for outer extensions. *Compositio Mathematica*, 20:29–52, 1969.

[74] Michael J. Fischer and Richard E. Ladner. Propositional dynamic logic of regular programs. *J. Comput. Syst. Sci.*, 1979.

[75] Massimo Franceschet and Maarten de Rijke. Model checking hybrid logics (with an application to semistructured data). *J. Applied Logic*, 4(3):279–304, 2006.

[76] Massimo Franceschet and Maarten de Rijke. Model checking for hybrid logics. In *Proceedings of the Workshop Methods for Modalities*, 2003.

[77] Nir Friedman and Joseph Y. Halpern. Modeling belief in dynamic systems, part I: foundations. *Artificial Intelligence*, 95(2):257–316, 1997.

[78] Drew Fudenberg and Jean Tirole. *Game Theory*. MIT Press, Cambridge, MA, 1991.

[79] Peter Gärdenfors. *Knowledge in Flux*. MIT Press, 1988.

[80] George Gargov and Salomon Passy. A note on boolean modal logic. In P. P. Petkov, editor, *Mathematical Logic*, pages 299–309, New York, London, 1990. Plenum Press. Proceedings of the Summer School and Conference on Mathematical Logic, honourably dedicated to the 90th Anniversary of Arend Heyting (1898–1980), Chaika, Bulgaria, 1988.

[81] George Gargov, Solomon Passy, and Tinko Tinchev. Modal environment for boolean speculations (preliminary report). In D. Skordev, editor, *Mathematical Logic and its Applications. Proceedings of the Summer School and Conference dedicated to the 80th Anniversary of Kurt Gödel*, pages 253–263. Plenum Press, Druzhba, 1987.

[82] John D. Geanakoplos and Heraklis M. Polemarchakis. We can't disagree forever. *Journal of Economic Theory*, 28(1):192–200, 1982.

[83] Jelle Gerbrandy. *Bisimulations on Planet Kripke*. PhD thesis, ILLC, Amsterdam, 1999.

[84] Edmund L. Gettier. Is justified true belief knowledge? *Analysis*, 23:121–123, 1963.

[85] Nina Gierasimczuk. Bridging learning theory and dynamic epistemic logic. *Synthese*, 169(2):371–384, 2009.

[86] Patrick Girard. *Modal Logic for Preference Change*. PhD thesis, Stanford, 2008.

[87] E. Mark Gold. Language identification in the limit. *Information and Control*, 10:447–474, 1967.

[88] Robert I. Goldblatt and S. K. Thomason. Axiomatic classes in propositional modal logic. In J. N. Crossley, editor, *Algebra and Logic: Papers 14th Summer Research Inst. of the Australian Math. Soc*, volume 450 of *LNM*, pages 163–173. Springer, Berlin, 1975.

[89] Valentin Goranko. Coalition games and alternating temporal logics. In *TARK '01*, pages 259–272, San Francisco, CA, USA, 2001. Morgan Kaufmann.

[90] Adam Grove. Two modellings for theory change. *Journal of Philosophical Logic*, 17:157–170, 1988.

[91] Yuri Gurevich. The classical decision problem. In *Perspectives in Mathematical Logic*. Springer, 1997.

[92] Joseph Y. Halpern. Substantive rationality and backward induction. *Games and Economic Behavior*, 37:2001, 1998.

[93] Joseph Y. Halpern and Yoram Moses. A guide to completeness and complexity for modal logics of knowledge and belief. *Artificial Intelligence*, 54 (3):319–379, 1992.

[94] Joseph Y. Halpern and Moshe Y. Vardi. The complexity of reasoning about knowledge and time. I. Lower bounds. *J. Comput. Syst. Sci.*, 38(1):195–237, 1989.

[95] Helle Hvid Hansen, Clemens Kupke, and Eric Pacuit. Bisimulations for neighbourhood structures. In *Proc. of 2nd Conference on Algebra and Coalgebra in CS*, LNCS, 2007.

[96] Sven Ove Hansson. Knowledge-level analysis of belief base operations. *Artif. Intell.*, 82(1–2):215–235, 1996.

[97] David Harel. Recurring dominoes: making the highly undecidable highly understandable. In *Topics in the theory of computation*, pages 51–71, New York, NY, USA, 1985. Elsevier.

[98] John C. Harsanyi. Games with incomplete information played by bayesian players. *Management Science*, 14:159–182, 320–334, 481–502, 1967.

[99] Xiangdong He, John F. Horty, and Eric Pacuit, editors. *Logic, Rationality, and Interaction, Second International Workshop, LORI 2009, Chongqing, China, October 8–11, 2009. Proceedings*, volume 5834 of *Lecture Notes in Computer Science*, 2009. Springer.

[100] Aviad Heifetz, editor. *TARK '09: Proceedings of the 12th Conference on Theoretical Aspects of Rationality and Knowledge*, New York, NY, USA, 2009. ACM.

[101] Edith Hemaspaandra. The price of universality. *Notre Dame Journal of Formal Logic*, 37(2):174–203, 1996.

[102] Jaakko Hintikka. *Knowledge and Belief: an introduction to the logic of the two notions.* Cornell University Press, 1962.

[103] Ian Hodkinson and Mark Reynolds. Temporal logic. In Patrick Blackburn, Johan van Benthem, and Frank Wolter, editors, *The Handbook of Modal Logic*, pages 655–720. Elsevier, 2006.

[104] Wiebe van der Hoek and John-Jules Ch. Meyer. Making some issues of implicit knowledge explicit. *Int. J. Found. Comput. Sci.*, 3(2):193–223, 1992.

[105] Wiebe van der Hoek and Marc Pauly. Modal logic for games and information. In Patrick Blackburn, Johan van Benthem, and Frank Wolter, editors, *The Handbook of Modal Logic*, pages 1077–1148. Elsevier, 2006.

[106] Tomohiro Hoshi. *Epistemic Dynamics and Protocol Information.* PhD thesis, Stanford, 2009.

[107] Sanjay Jain, Daniel Osherson, James S. Royer, and Arun Sharma. *Systems that Learn: An Introduction to Learning Theory.* MIT Press, Cambridge, Massachusetts, 2nd edition, 1999.

[108] Richard C. Jeffrey. *The Logic of Decision.* University of Chicago Press, 2nd edition, 1983.

[109] Laurent Keiff. *Le Pluralisme Dialogique*. PhD thesis, Université Charles de Gaulle, Lille, 2007.

[110] Dexter Kozen and Rohit Parikh. An elementary proof of the completeness of PDL. *Theoretical Computer Science*, 14:113–118, 1981.

[111] David M. Kreps, Paul Milgrom, John Roberts, and Robert Wilson. Rational cooperation in the finitely repeated prisoners' dilemma. *Journal of Economic Theory*, 27(2):245–252, 1982.

[112] Lena Kurzen. Logics for Cooperation, Actions and Preferences. Master's thesis, Universiteit van Amsterdam, Netherlands, 2007.

[113] Richard E. Ladner. The computational complexity of provability in systems of modal propositional logic. *SIAM J. Comput.*, 6(3):467–480, 1977.

[114] Martin Lange. Model checking PDL with all extras. *J. Applied Logic*, 4(1): 39–49, 2006.

[115] Steffen Lange, Rolf Wiehagen, and Thomas Zeugmann. Learning by erasing. In *Proc. 7th Int. Workshop on Algorithmic Learning Theory*, Lecture Notes in Artificial Intelligence, pages 228–241. Springer-Verlag, 1996.

[116] Isaac Levi. Subjunctives, dispositions and chances. *Synthese*, 34:423–455, 1977.

[117] David K. Lewis. Elusive knowledge. *Australasian Journal of Philosophy*, 74(4):549–567, 1996. 418-446.

[118] David K. Lewis. *Counterfactuals*. Harvard University Press, Cambridge, 1973.

[119] Fenrong Liu. *Changing for the Better: Preference Dynamics and Agent Diversity*. PhD dissertation, ILLC Amsterdam, 2008.

[120] Carsten Lutz and Ulrike Sattler. The complexity of reasoning with boolean modal logics. In Wolter, Wansing, de Rijke, and Zakharyaschev, editors, *AiML*, pages 329–348. WS, 2000. ISBN 981-238-179-1.

[121] Lucie Ménager. *Communication, common knowledge, and consensus*. PhD thesis, Universite Paris I Pantheon-Sorbonne, 2006.

[122] Ron van der Meyden and Ka-shu Wong. Complete axiomatizations for reasoning about knowledge and branching time. *Studia Logica*, 75(1):93–123, 2003.

[123] Yasuhito Mukouchi. Characterization of finite identification. In *AII '92: Proceedings of the International Workshop on Analogical and Inductive Inference*, pages 260–267, London, UK, 1992. Springer-Verlag.

[124] Roger B. Myerson. *Game Theory: Analysis of Conflict*. Harvard University Press, 1997.

[125] Piergiorgio Odifreddi. *Classical Recursion Theory*. Number 125 in Studies in Logic and the Foundations of Mathematics. North-Holland, 1989.

[126] Martin J. Osborne. *An Introduction to Game Theory*. Oxford University Press, USA, 2004.

[127] Martin J. Osborne and Ariel Rubinstein. *A course in game theory*. MIT Press, Cambridge, MA, 1994.

[128] Eric Pacuit. Some comments on history based structures. *Journal of Applied Logic*, 5(4):613–624, 2007.

[129] Christos M. Papadimitriou. *Computational complexity*. Addison-Wesley, MA, 1994.

[130] Rohit Parikh and Ramaswamy Ramanujam. A knowledge based semantics of messages. *Journal of Logic, Language and Information*, 12(4):453–467, 2003.

[131] Marc Pauly. A modal logic for coalitional power in games. *Journal of Logic and Computation*, 12(1):149–166, 2002.

[132] Marc Pauly and Rohit Parikh. Game logic — an overview. *Studia Logica*, 75(2):165–182, 2003.

[133] David G. Pearce. Rationalizable strategic behavior and the problem of perfection. *Econometrica*, 52(4):1029–1050, 1984.

[134] Jan A. Plaza. Logics of public communications. In M.L. Emrich, M.S. Pfeifer, M. Hadzikadic, and Z.W. Ras, editors, *Proceedings of the Fourth International Symposium on Methodologies for Intelligent Systems: Poster Session Program*, pages 201–216. Oak Ridge National Laboratory, 1989.

[135] Antonio Quesada. On expressing maximum information in extensive games. *Mathematical Social Sciences*, 42(2):161–167, 2001.

[136] Shahid Rahman and Laurent Keiff. On how to be a dialogician. In Daniel Vanderveken, editor, *Logic, Thought and Action*, Logic, Epistemology and the Unity of Science, pages 359–408. Springer Verlag, Dordrecht, 2004.

[137] John Rawls. *A theory of justice*. Harvard University Press, Cambridge, 1971.

[138] Philip J. Reny. Backward induction, normal form perfection and explicable equilibria. *Econometrica*, 60(3):627–49, 1992.

[139] Philip J. Reny. Common belief and the theory of games with perfect information. *Journal of Economic Theory*, 59(2):257–274, April 1993.

[140] Stéphane Le Roux, Pierre Lescanne, and René Vestergaard. Conversion/preference games. *CoRR*, abs/0811.0071, 2008.

[141] Henrik Sahlqvist. Completeness and correspondence in the first and second order semantics for modal logic. In Stig Kanger, editor, *Proceedings of the Third Scandinavian Logic Symposium*, pages 110–143. Uppsala, 1973.

[142] Dov Samet. Agreeing to disagree: The non-probabilistic case. *Games and Economic Behavior*, In Press.

[143] Krister Segerberg. Bringing it about. *Journal of Philosophical Logic*, 18(4): 327–347, 1989.

[144] Edith Spaan. *Complexity of modal logics*. PhD thesis, ILLC Amsterdam, 1993.

[145] Michael Spence. *Market signaling: informational transfer in hiring and related screening processes*. Harvard University Press, Cambridge, 1974.

[146] Robert Stalnaker. Belief revision in games: forward and backward induction. *Mathematical Social Sciences*, 36(1):31–56, 1998.

[147] Robert Stalnaker. Knowledge, belief and counterfactual reasoning in games. *Economics and Philosophy*, 12(02):133–163, 1996.

[148] Robert Stalnaker. A theory of conditionals. In Robert Stalnaker, William L. Harper, and Glenn Pearce, editors, *Ifs: Conditionals, Belief, Decision, Chance and Time*. D. Reidel, Dordrecht, 1981.

[149] Colin Stirling. Modal and temporal logics. In S. Abramsky, D.M. Gabbay, and T.S.E. Maibaum, editors, *Handbook of logic in computer science (vol. 2) — Background: Computational structures*, pages 478–563. Oxford University Press, New York, NY, 1992.

[150] 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.

[151] Moshe Y. Vardi. The complexity of relational query languages (extended abstract). In *STOC '82: Proceedings of the fourteenth annual ACM symposium on Theory of computing*, pages 137–146, New York, NY, USA, 1982. ACM.

[152] Jonathan A. Zvesper. *Playing with information.* PhD thesis, ILLC Amsterdam, 2009.

Samenvatting

Het algemene analytische project waartoe deze dissertatie bijdraagt is het modelleren van redeneren door rationele actoren in interactieve situaties, en het identificeren van de logica hiervan. De grenzen van dit project doorkruisen de economie, de informatica en de filosofie, in het bijzonder de volgende deelgebieden. Interactieve epistemologie bestudeert interactief redeneren: hoe hebben en verwerven actoren kennis en hoe redeneren zij over de kennis van anderen. Formele leertheorie bestudeert de voorwaarden waaronder een actor stabiele kennis kan bereiken, en uit een stroom van data een juiste hypothese kan vinden. Epistemische speltheorie bestudeert hoe rationele actoren beslissingen nemen tot handelen op grond van hun kennis in strategische interactieve situaties. In al deze systemen staan interactieve vormen van kennis en geloof centraal, en hun evolutie door informatie-processen in de tijd.

In deze dissertatie worden al deze onderwerpen verbonden in één logisch systeem. Daarmee bevinden we ons op het grensvlak van twee grote paradigma's die verandering van kennis beschrijven: de temporele en de dynamische logica. We vergelijken deze, en combineren ze vervolgens: eerst op een structureel-semantisch niveau, vervolgens ook syntactisch. Het aldus verkregen systeem passen we toe in de analyse van wat er gebeurt met de kennis van actoren gedurende een tijdspanne waarin zij communiceren, leren, op elkaar reageren, en inductief of strategisch redeneren.

Hoofdstuk 2 bepaalt de algemene structurele eigenschappen van actoren die kennis en geloof herzien op grond van informatie door de tijd heen, en in hoofdstuk 3 worden hiervoor dan algemene logische bewijsprincipes gegeven in een epistemisch-temporele logica. De analyse verloopt technisch via semantische representatie-stellingen, alsmede een volledigheidsbewijs voor een tijdslogica van kennisverandering voor modellen met informatie-protocollen. De volgende hoofdstukken bestuderen concrete informatieprocessen tegen deze algemene achtergrond. Hoofstuk 4 onderzoekt hoe actoren door herhaalde aankondigingen van wat zij geloven op den duur tot overeenstemming kunnen komen, en identificeert

gemeenschappelijk geloof in 'posteriors' in geschikte structuren als de voornaamste voorwaarde. Tevens definiëren we statische en dynamische logica's waarmee we over processen van overeenstemming kunnen redeneren, en vinden enkele belangrijke semantische en bewijstheoretische eigenschappen. Hoofdstuk 5 onderzoekt logische principes van inductief leren door de tijd heen, en in het bijzonder van de centrale leertheoretische notie van eindige identificeerbaarheid. De vorm die dit aanneemt is een reductie van eindige identificeerbaarheid tot een semantische evaluatietaak voor epistemisch-temporele logica. Voorts worden enkele representatie resultaten gegeven voor de bijbehorende modellen. Hoofdstuk 6 past het door ons ontwikkelde dynamisch-temporele perspectief toe op de bouwstenen van strategisch redeneren met onvolledige informatie: oplossingsalgorithmen voor spelen, rationaliteit, speltheoretisch evenwicht, en verwachtingen van spelers. Hierbij wordt het belang van dynamische kennislogica voor de epistemische grondslagen van de speltheorie geïllustreerd aan de hand van vele concrete situaties. Hoofdstuk 7 rondt onze benadering af met twee verdere belangrijke thema's: voorkeuren van actoren, en groepshandelen (in het bijzonder, de vermogens van coalities). We onderzoeken vanuit ons logisch perspectief een groot aantal voorgestelde noties uit de sociale keuzetheorie en non-coöperatieve speltheorie. Dit leidt tot vele resultaten over logische definiëerbaarheid, semantische invariantie en computationele complexiteit.

Abstract

Modeling rational agents' reasoning in interactive contexts and identifying its logic is the general analytic project to which this dissertation contributes. The borders of this project run through economics, computer science and philosophy. It includes several theoretical lines that we are connecting. Interactive epistemology is the study of interactive reasoning: how agents entertain beliefs and reason about the beliefs of other agents. Formal learning theory is the study of the conditions under which agents can reach stable beliefs or identify a correct hypothesis from a stream of data. Epistemic game theory is a theory of how rational agents would make decisions based on their beliefs in strategic interactive situations. In all these systems, beliefs, interactive beliefs, and their evolution as informational processes unfold are at stake.

This dissertation connects these themes by developing one single logical framework. For this purpose, we are operating at the interface of two major logics of belief change: the temporal approach and the dynamic approach. Concretely, we connect and merge the two families of logics, first at a structural semantic level and then at a syntactic one. Subsequently, we apply the resulting system to analyze what happens to agents' beliefs over time when agents communicate, learn, interact, and reason interactively, inductively, or strategically.

Chapter 2 identifies the main structural properties of belief revising agents over time, and Chapter 3 then formulates their main logical proof principles. This chiefly takes the form of semantic representation theorems, plus a completeness theorem for changing beliefs in a temporal logic that admits protocols. Chapter 4 identifies common belief of posteriors in suitable structures as a key sufficient condition for agents to agree, and iterated announcement of beliefs as a major way of reaching agreement. We also determine the right family of static and dynamic logics to reason about agreement, and find agreement results, invariance results, and concrete syntactic proofs of agreement results. Chapter 5 investigates the logical principles behind inductive learning and in particular behind the key notion of finite identifiability. This takes the form of a reduction of the problem

of finite identifiability to a problem of model-checking for an epistemic temporal logic, plus further representation results. Chapter 6 takes the dynamic-temporal logical viewpoint to the building blocks of strategic reasoning: solution algorithms, rationality, equilibrium, and expectations, discussing the importance of belief change for the epistemic foundations of game theory. We are giving many concrete scenarios sketching a bigger picture. Chapter 7 completes the whole approach with two further key aspects of agency: preferences, and coalitional powers. We explore the logical expressive power demanded by notions imported in this area from social choice theory and cooperative and non-cooperative game theory, in terms of modal invariance and definability.

Résumé

Modéliser la façon dont des agents rationnels raisonnent en contextes interactifs puis identifier la logique de ce raisonnement, est le projet analytique global à laquelle cette thèse contribue. Les frontières de ce projet traversent les sciences économiques, les sciences de l'information et la philosophie. Nous connectons certaines lignes de recherche importantes qui s'inscrivent dans ce projet. Ces lignes sont l'épistémologie interactive, la théorie formelle de l'apprentissage et la théorie des jeux épistémique. L'épistémologie interactive étudie le raisonnement interactif, la façon dont des agents entretiennent des croyances et raisonnent à propos des croyances d'autres agents. La théorie formelle de l'apprentissage s'intéresse aux conditions auxquelles des agents peuvent parvenir à avoir des croyances stables ou identifier une hypothèse correcte à partir d'une séquence de données. La théorie des jeux épistémique a pour objet la façon dont des agents rationnels prendraient des décisions fondées sur leurs croyances dans des situations interactives stratégiques. Dans tous ces systèmes, les croyances, les croyances interactives, et leur évolution à mesure que les agents reçoivent et échangent de nouvelles informations, occupent une place centrale.

Cette thèse relie les thèmes précédents en développant un cadre logique unique. Pour ce faire, nous travaillons à l'interface de deux approches logiques importantes du changement de croyances: l'approche temporelle et l'approche dynamique. Plus précisément, nous relions et fusionnons les deux familles de logiques, tout d'abord à un niveau sémantique structurel, puis à un niveau syntaxique. Nous appliquons le système ainsi obtenu à l'analyse de l'évolution temporelle des croyances d'agents lorsqu'ils communiquent, apprennent, interagissent et raisonnent de façon interactive, inductive ou stratégique.

Le chapitre 2 identifie les propriétés doxastiques temporelles structurelles qui caractérisent les agents révisant leurs croyances dynamiquement, et le chapitre 3 en formule ensuite les principes logiques centraux. Cela prend principalement la forme de théorèmes de représentation sémantiques, et d'un théorème de complétude pour une logique du changement de croyances qui autorise les pro-

tocoles. Le chapitre 4 identifie la croyance commune des croyances postérieures dans des structures qualitatives comme une condition-clé suffisante pour garantir que des agents s'accordent, et l'annonce itérée de croyances comme un moyen important d'y parvenir. Nous déterminons également la famille de logiques statiques et dynamiques adéquate pour raisonner à propos de problèmes d'accord. Nous trouvons de tels résultats d'accord, mais aussi des résultats d'invariance, et des preuves syntaxiques pour des théorèmes d'accord. Le chapitre 5 s'intéresse aux principes logiques du raisonnement inductif et à ceux qui soutiennent la notion d'identifiabilité finie. Cela prend la forme d'une réduction du problème d'identifiabilité finie à un problème de vérification (*model checking*) pour une logique temporelle épistémique, et de résultats de représentation. Le chapitre 6 oriente le point de vue logique dynamique-temporel sur les concepts élémentaires du raisonnement stratégique : algorithmes de solution, rationalité, équilibre, et anticipations, discutant l'importance du changement de croyances pour les fondations épistémiques des jeux. Nous donnons de nombreux scénarios concrets esquissant une image plus globale. Le chapitre 7 enrichit l'approche en considérant deux autres aspects clés complémentaires de l'interaction: les préférences et le pouvoir des coalitions. Nous explorons le pouvoir expressif requis par des notions importées dans le champ logique à partir du choix social et de la théorie des jeux coopérative et non-coopérative, en termes d'invariance modale et de définissabilité.

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

ILLC DS-2010-02: **Jonathan Zvesper**
Playing with Information